Unbinned and High-dimensional Unfolding with Machine Learning

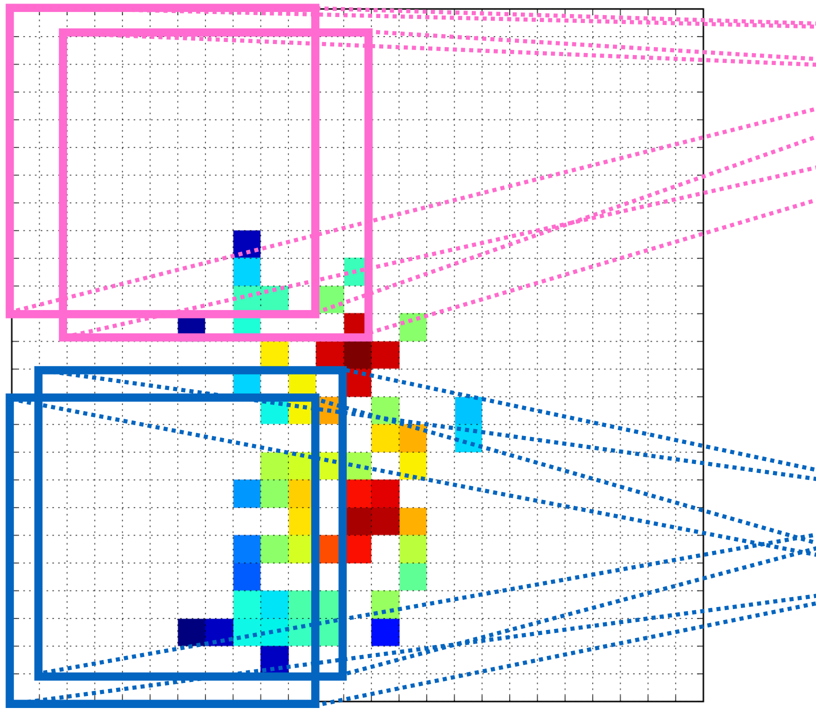
Benjamin Nachman

Lawrence Berkeley National Laboratory

bpnachman.com
bpnachman@lbl.gov

@bpnachman

bnachman



BERKELEY LAB

BIDS
BERKELEY INSTITUTE
FOR DATA SCIENCE

PHYSTAT
Seminar
March 2023

Unfolding

Deconvolution ("unfolding"):
correcting for detector effects

Unfolding

Deconvolution ("unfolding"):
correcting for detector effects

Key aspect of all cross section measurements, across particle/nuclear/astro physics (!)

Unfolding

Proton-Proton

Nucleus-Nucleus

Electron-Proton

Neutrino-Nucleus

Electron-Positron

Cosmic Rays

Particle/Nuclear/Astro Physics Experiments

Deconvolution (“unfolding”): correcting for detector effects

Key aspect of **all** cross section measurements, across particle/nuclear/astro physics (!)

Unfolding

Proton-Proton

Nucleus-Nucleus

Electron-Proton

Neutrino-Nucleus

Electron-Positron

Cosmic Rays

Particle/Nuclear/Astro Physics Experiments

Deconvolution ("unfolding"): correcting for detector effects

Key aspect of **all** cross section measurements, across particle/ nuclear/astro physics (!)

Why "unfold" instead of "fold"?

Unfolding is ill-posed, BUT only way to compare different experiments and to compare with non fully exclusive predictions. Data also survive much longer.

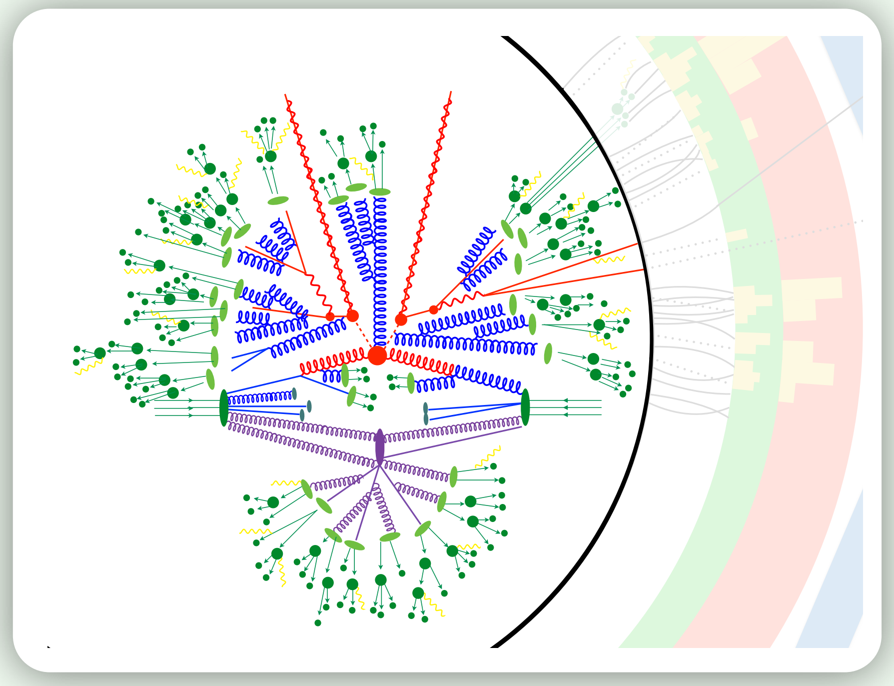
The Unfolding Challenge

2203.16722

The Unfolding Challenge

Particle Level

Want this



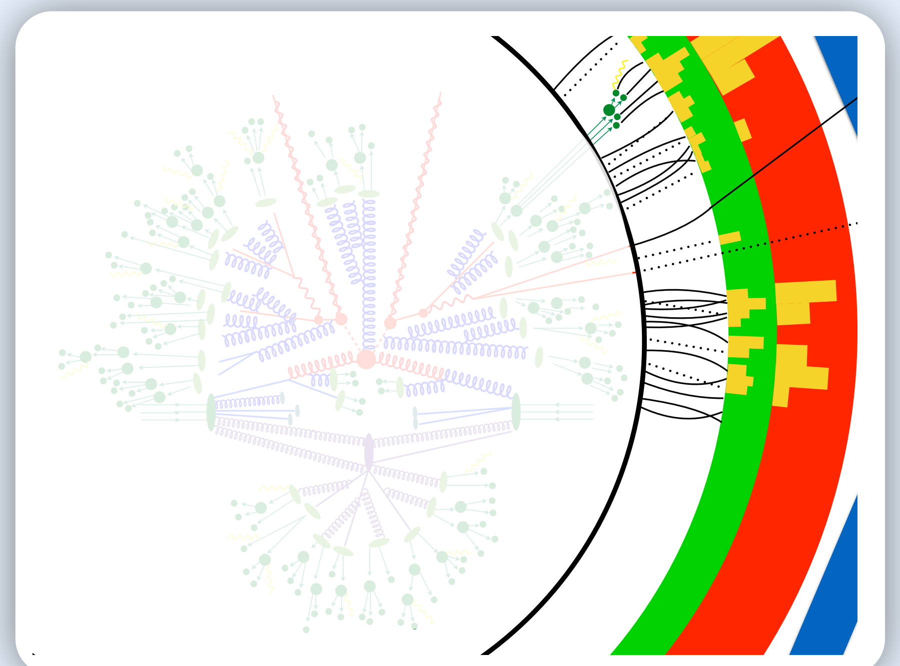
2203.16722

Detector Level

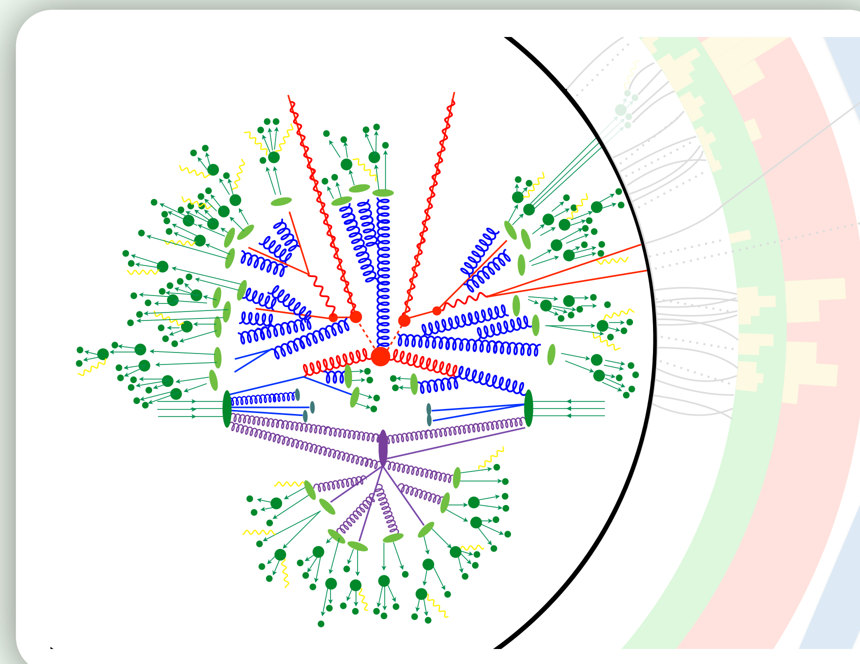
The Unfolding Challenge

Particle Level

Measure this



Want this



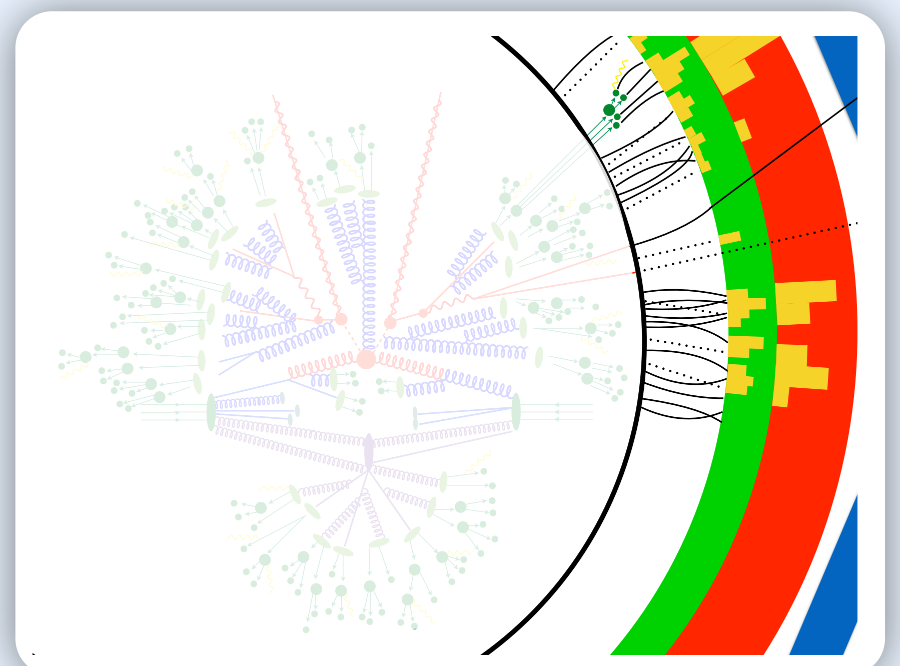
2203.16722

Detector Level

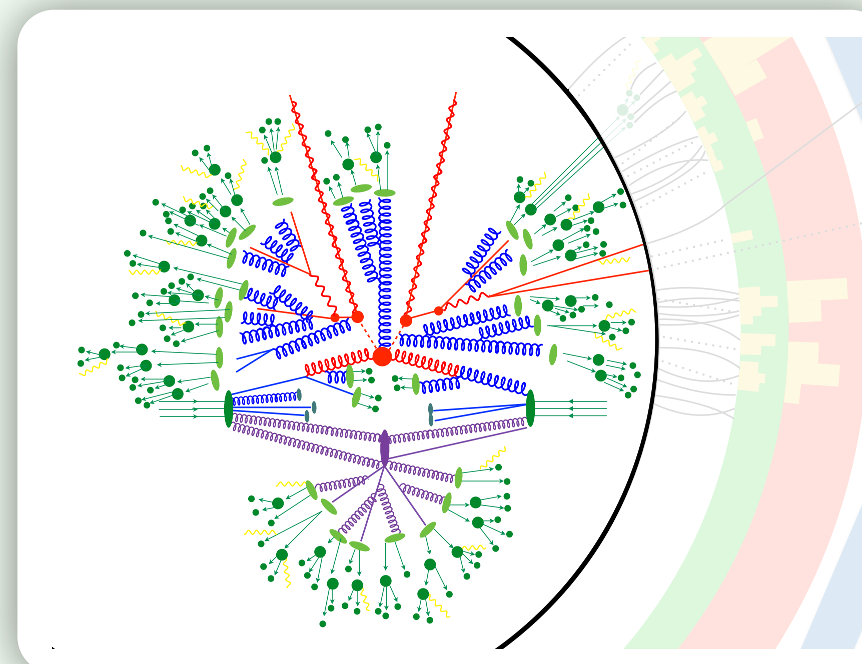
The Unfolding Challenge

Particle Level

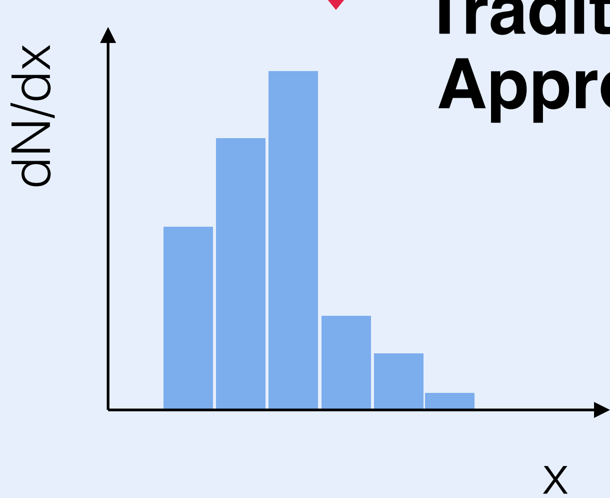
Measure this



Want this

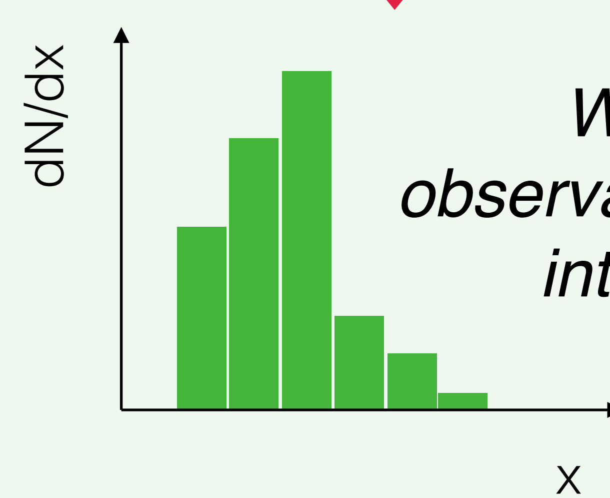


Target observable



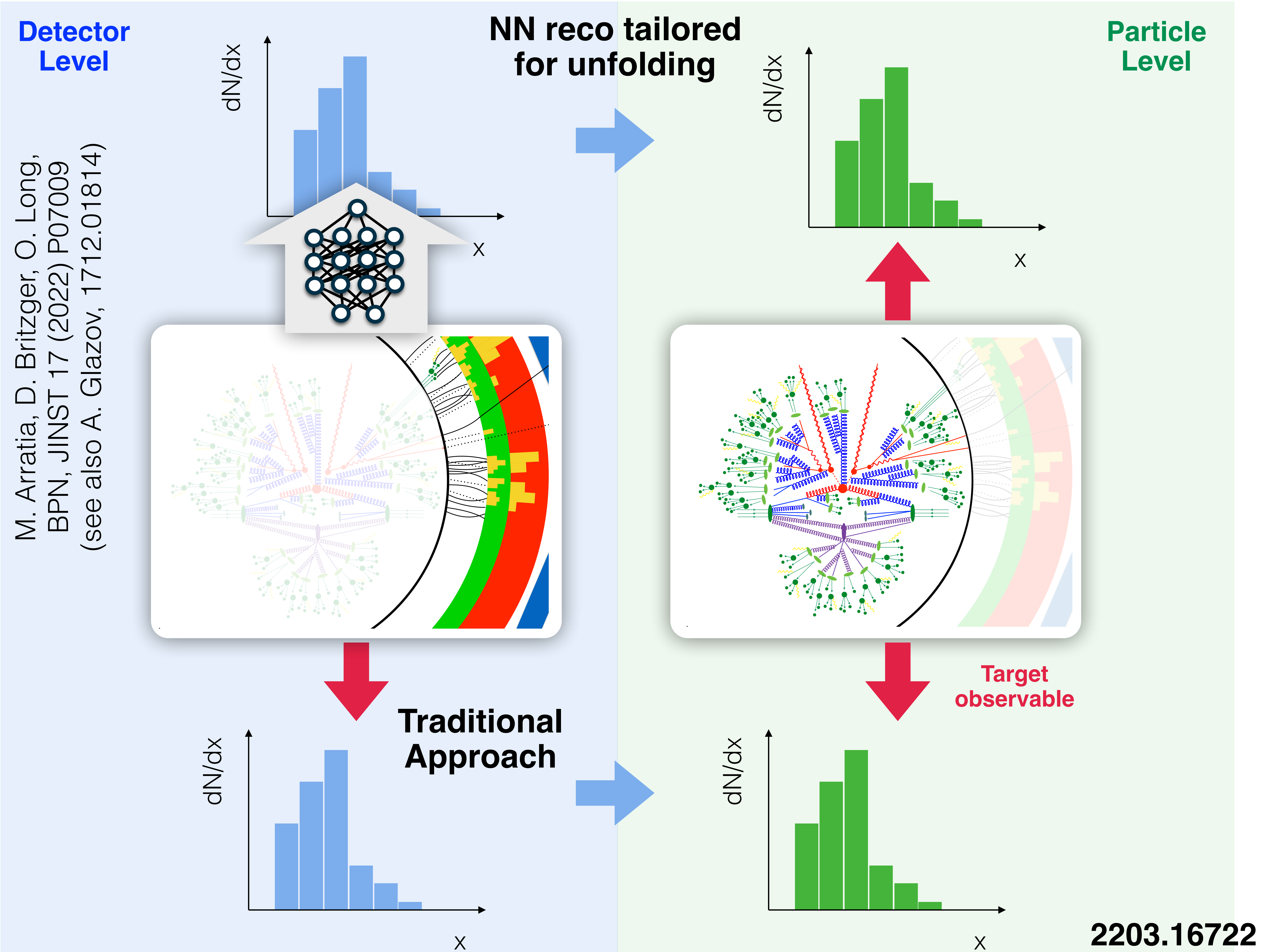
Traditional Approach

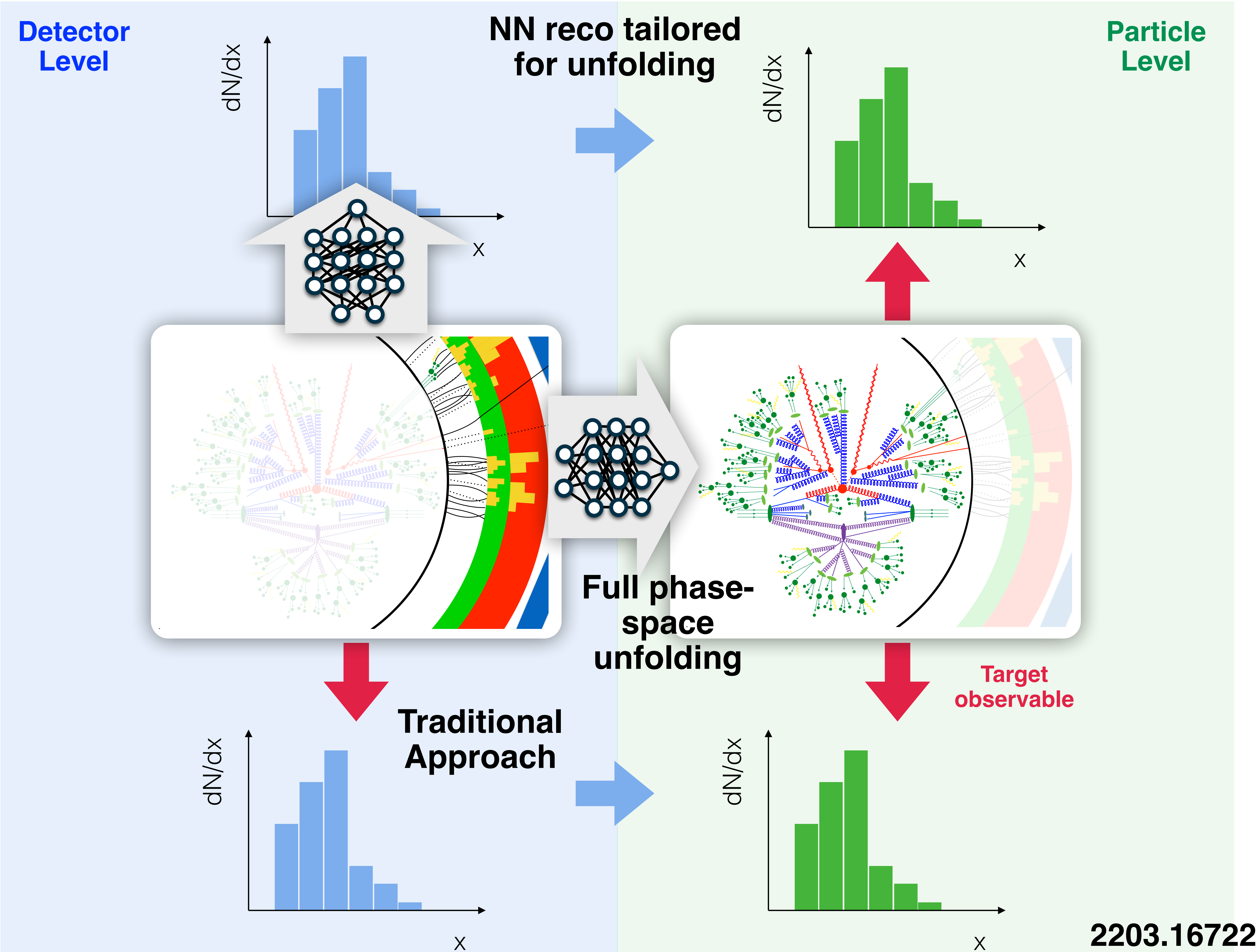
$$\hat{T} \approx R^{-1}M$$

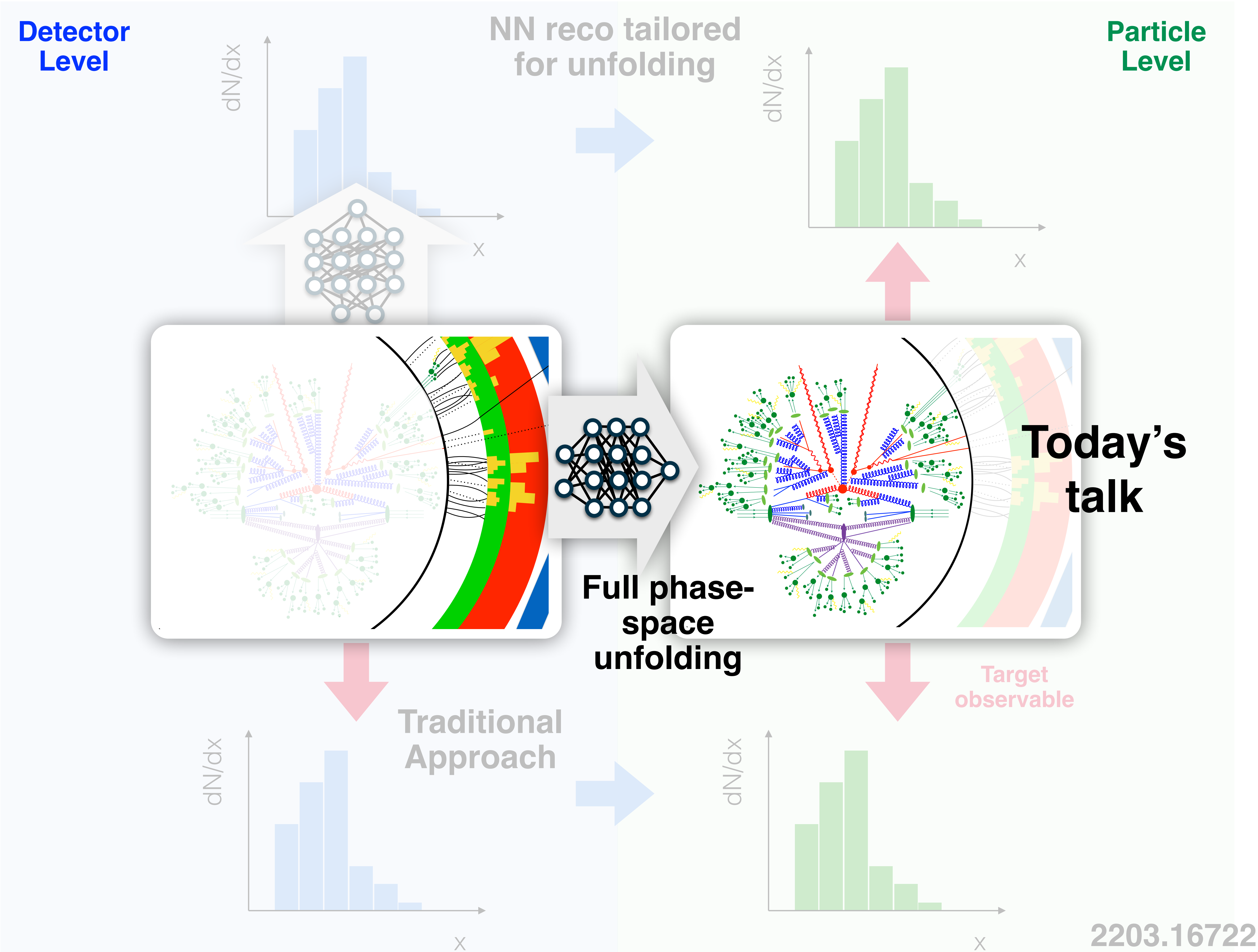


We pick O(1) observables & discretize into O(10) bins

2203.16722


Detector Level
NN reco tailored for unfolding
Particle Level
M. Arratia, D. Britzger, O. Long, BPN, JINST 17 (2022) P07009 (see also A. Glazov, 1712.01814)
dN/dx
x
Target observable
Traditional Approach
2203.16722


Detector Level
NN reco tailored for unfolding
Particle Level
dN/dx
x
dN/dx
x
Full phase-space unfolding
Target observable
Traditional Approach
dN/dx
x
dN/dx
x
2203.16722



Detector Level
NN reco tailored for unfolding
Particle Level
dN/dx
x
dN/dx
x
Today's talk
Full phase-space unfolding
Target observable
Traditional Approach
dN/dx
x
dN/dx
x
2203.16722

Why unbinned (+high-dimensional)?

For a community white paper, see JINST 17 (2022) P01024, 2109.13243

Why unbinned (+high-dimensional)?

Inference-Aware Binning

Optimal binning depends on downstream task. Not possible with current setup.

What about moments?
(see also K. Desai, BPN, J. Thaler, [paper])

For a community white paper, see JINST 17 (2022) P01024, 2109.13243

Why unbinned (+high-dimensional)?

Inference-Aware Binning

Optimal binning depends on downstream task. Not possible with current setup.

What about moments?
(see also K. Desai, BPN, J. Thaler, [paper])

Derivative Measurements

With binned measurements, essentially impossible to re-use results for a function of the phase space.

For a community white paper, see JINST 17 (2022) P01024, 2109.13243

Why unbinned (+high-dimensional)?

Inference-Aware Binning

Optimal binning depends on downstream task. Not possible with current setup.

What about moments?
(see also K. Desai, BPN, J. Thaler, [paper])

Derivative Measurements

With binned measurements, essentially impossible to re-use results for a function of the phase space.

Higher Dimensions

Some phenomena can't be probed in a few dimensions.

What about observables that are not per-event?

For a community white paper, see JINST 17 (2022) P01024, 2109.13243

Landscape of Methods

Classifier-Based Methods

Learn (unfolded) data likelihood ratio w.r.t. simulation

For references, see JINST 17 (2022) P01024, 2109.13243 (and papers that cite it!)

Landscape of Methods

Classifier-Based Methods

Learn (unfolded) data likelihood ratio w.r.t. simulation

Density-Based Methods

Learn (unfolded) data probably density implicitly or explicitly.

For references, see JINST 17 (2022) P01024, 2109.13243 (and papers that cite it!)

Landscape of Methods

Classifier-Based Methods

Learn (unfolded) data likelihood ratio w.r.t. simulation

I'll focus here today because:

Learn a small correction (start close to the right answer)

&

Prior independent (if maximum likelihood)

Density-Based Methods

Learn (unfolded) data probably density implicitly or explicitly.

For references, see JINST 17 (2022) P01024, 2109.13243 (and papers that cite it!)

Landscape of Methods

Classifier-Based Methods

Learn (unfolded) data likelihood ratio w.r.t. simulation

I'll focus here today because:

Learn a small correction (start close to the right answer)

&

Prior independent (if maximum likelihood)

Density-Based Methods

Learn (unfolded) data probably density implicitly or explicitly.

I won't talk about these at all, but there has been a lot of work with GANs, VAEs, NFs, …

GANs: K. Datta, D. Kar, D. Roy, 1806.00433; M. Bellagente, A. Butter, G. Kasieczka, T. Plehn, R. Winterhalder, SciPost Phys. 8 (2020) 070, …

VAEs: J. Howard, S. Mandt, D. Whiteson, Y. Yang, Sci. Rep. 12 (2022) 7567, …

NFs: M. Bellagente et al., SciPost Phys. 9 (2020) 074; M. Vandegar, M. Kagan, A. Wehenkel, G. Louppe, PMLR 11 (2021) 2107; M. Backes, A. Butter, M. Dunford, B. Malaescu, 2212.08674, …

For references, see JINST 17 (2022) P01024, 2109.13243 (and papers that cite it!)

Landscape of Methods

Classifier-Based Methods

Learn (unfolded) data likelihood ratio w.r.t. simulation

I'll focus here today because:

Learn a small correction (start close to the right answer)

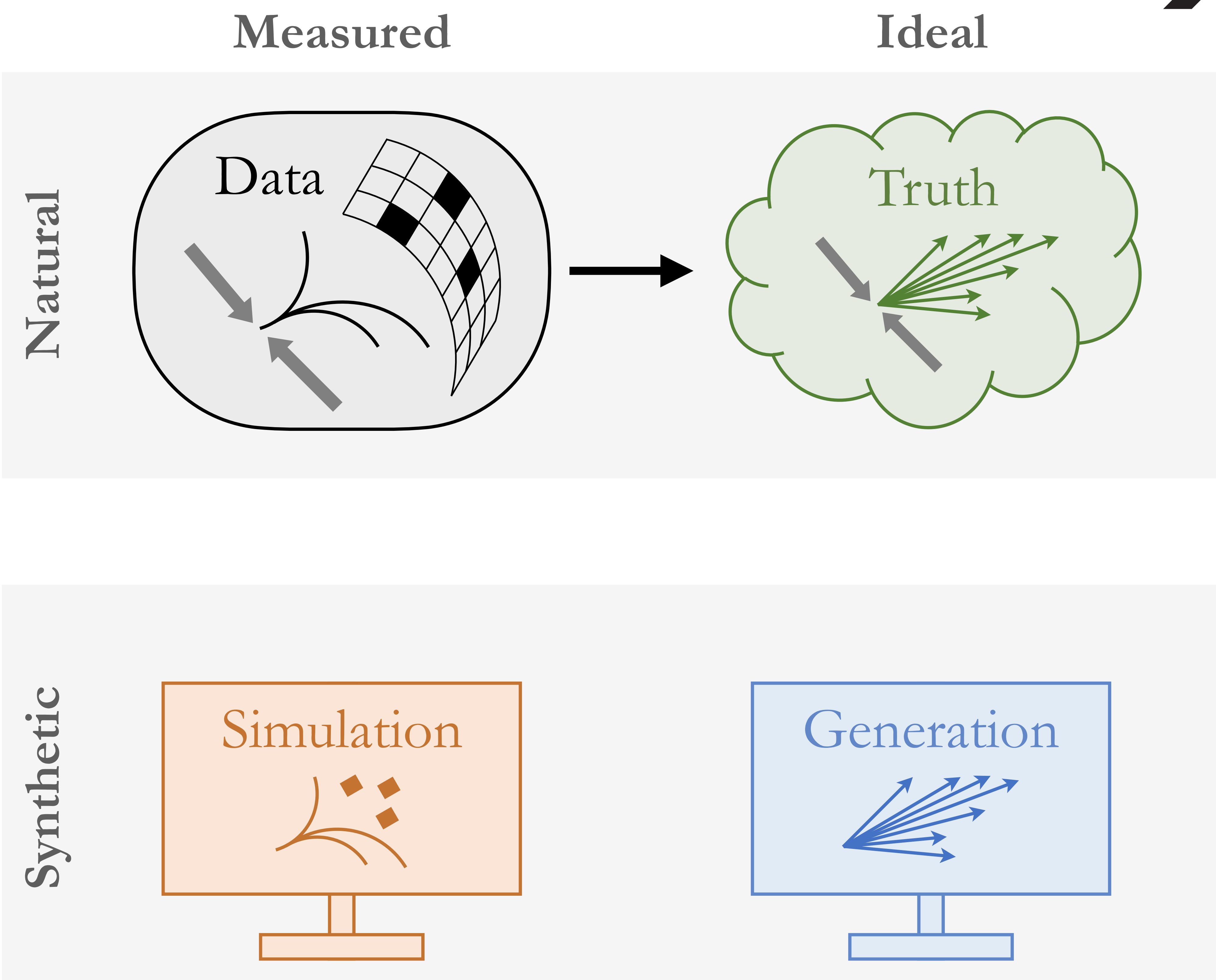
&

Prior independent (if maximum likelihood)

My focus will be on a method called **OmniFold**.

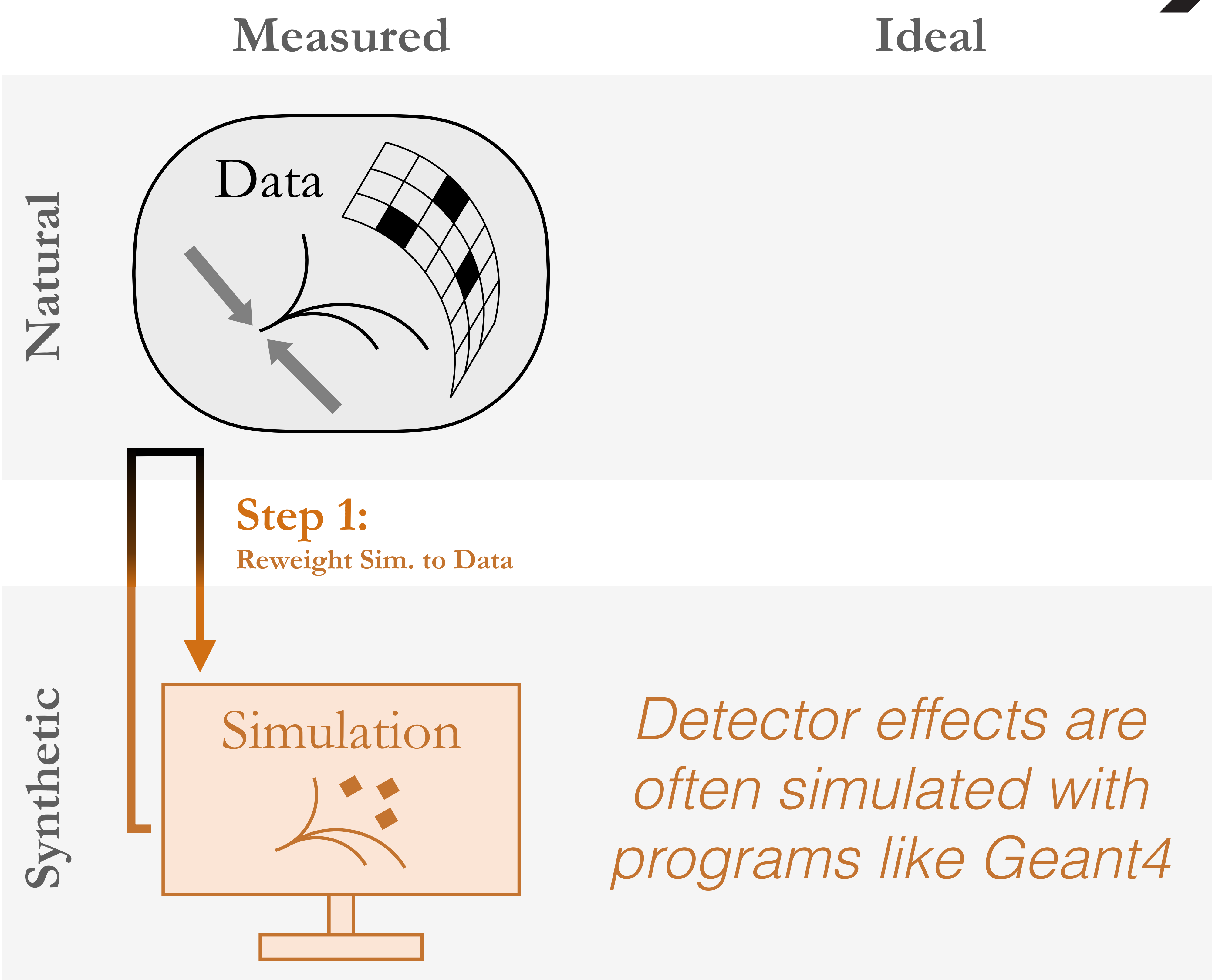
For references, see JINST 17 (2022) P01024, 2109.13243 (and papers that cite it!)

Unfold by iterating: OmniFold



A. Andreassen, P. Komiske, E. Metodiev,
BPN, J. Thaler, PRL 124 (2020) 182001

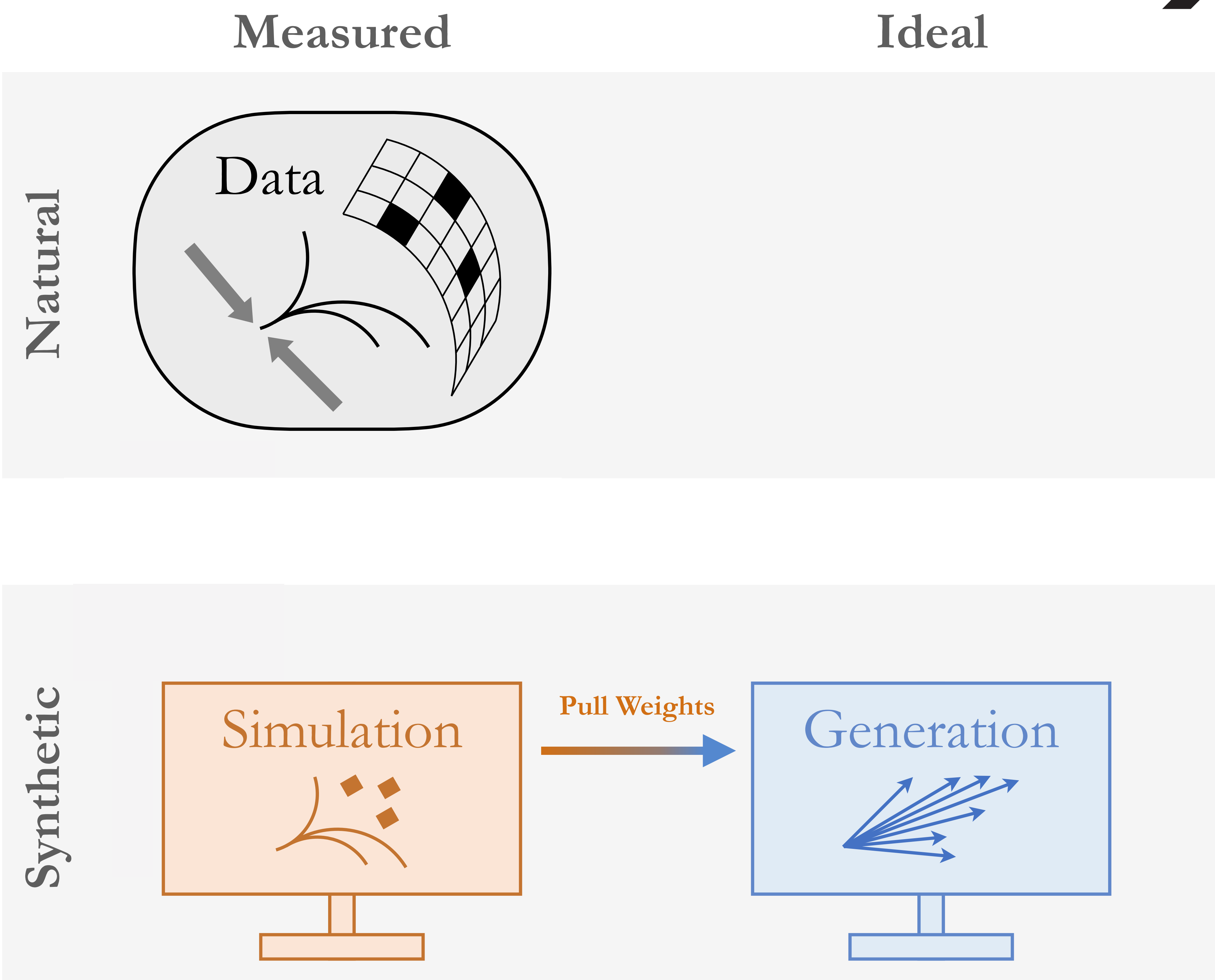
Unfold by iterating: OmniFold



Detector effects are often simulated with programs like Geant4

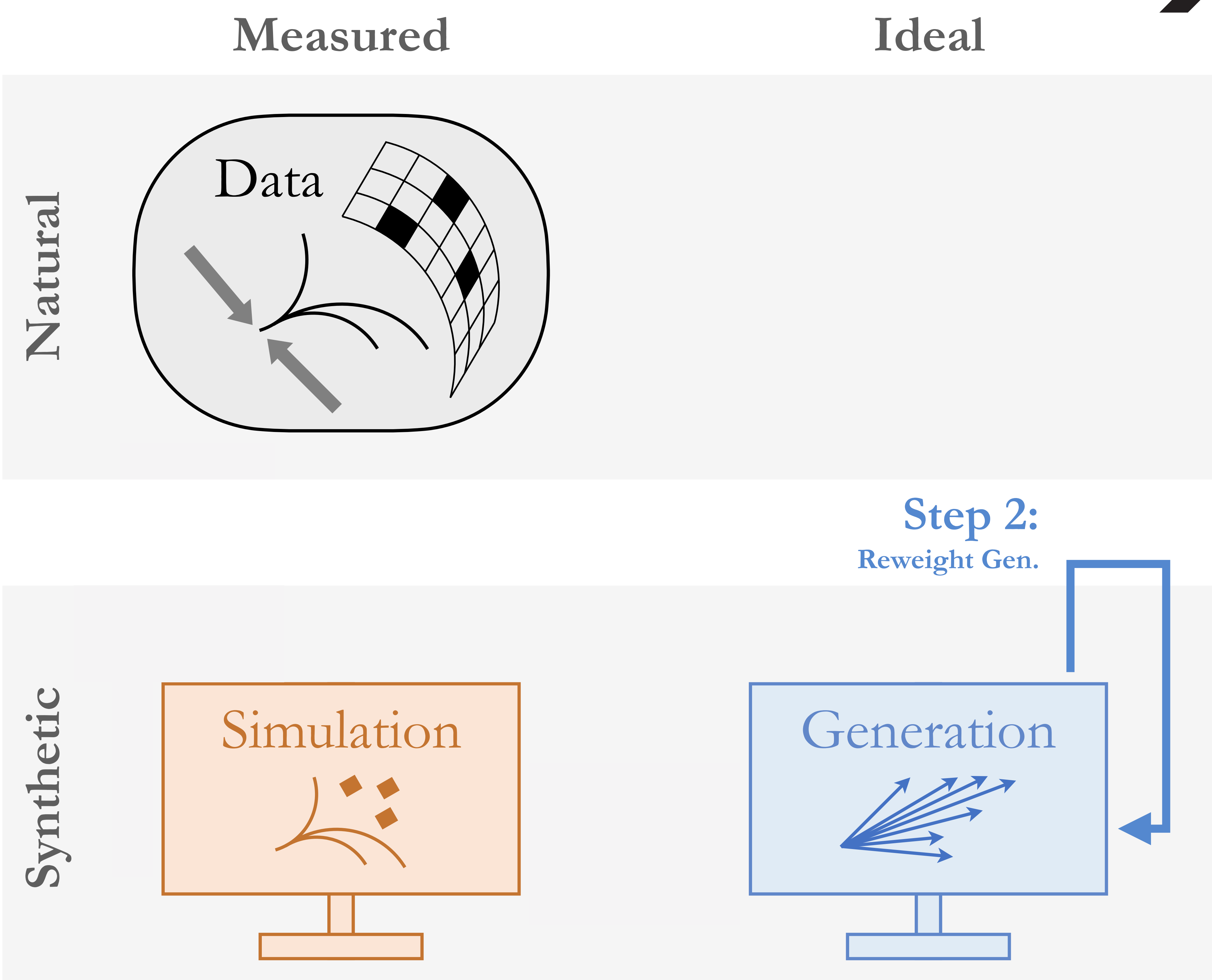
A. Andreassen, P. Komiske, E. Metodiev, BPN, J. Thaler, PRL 124 (2020) 182001

Unfold by iterating: OmniFold



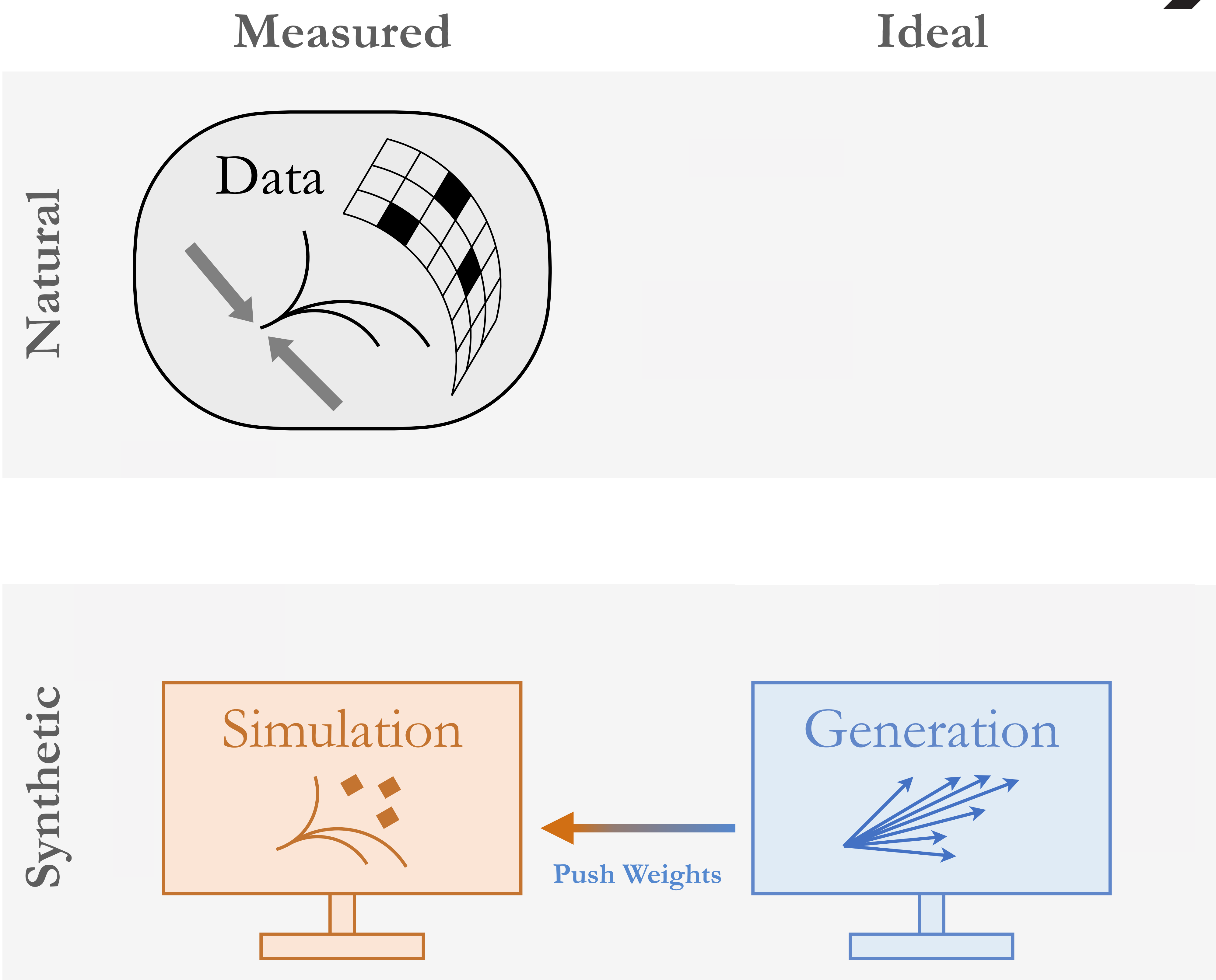
A. Andreassen, P. Komiske, E. Metodiev, BPN, J. Thaler, PRL 124 (2020) 182001

Unfold by iterating: OmniFold



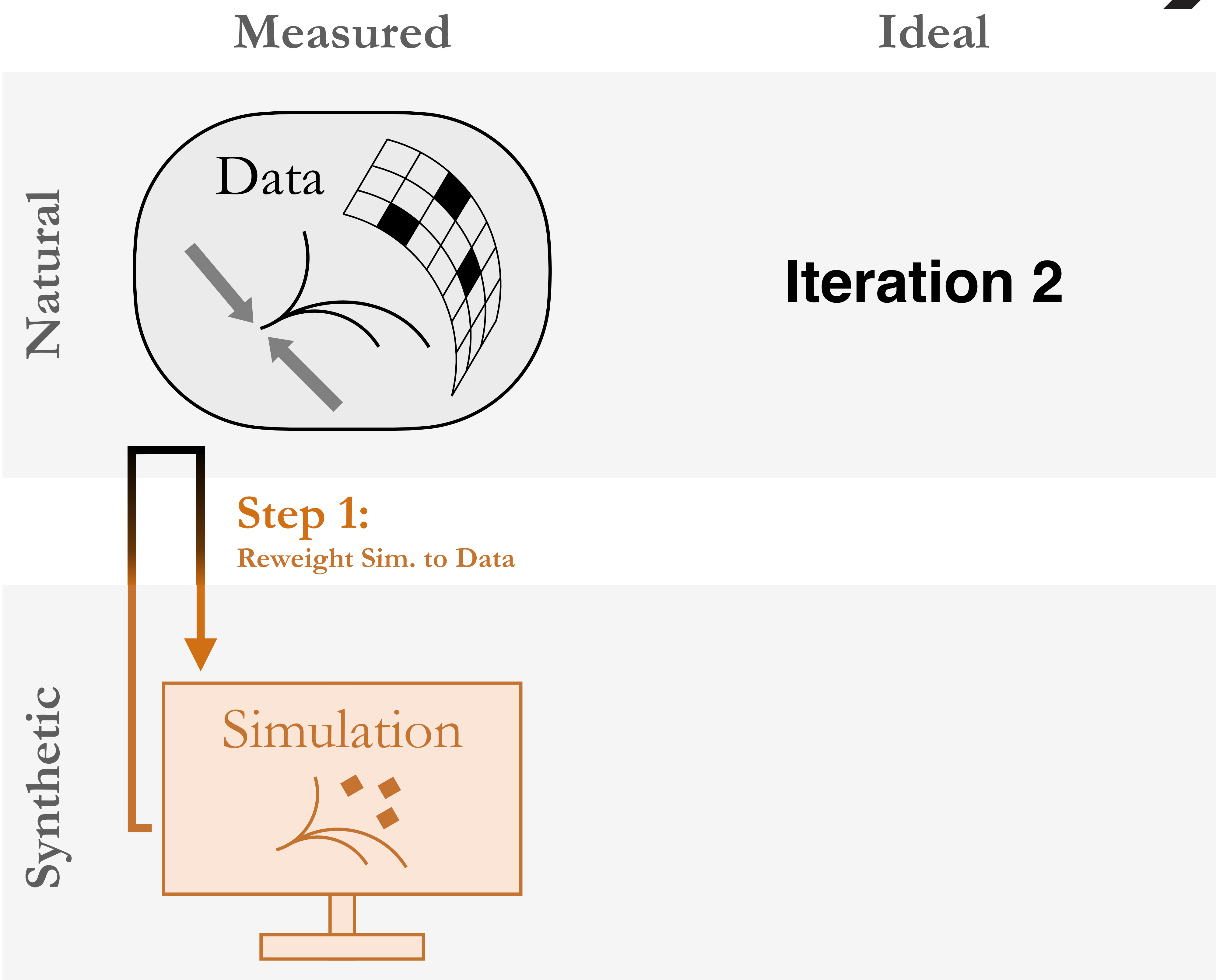
A. Andreassen, P. Komiske, E. Metodiev, BPN, J. Thaler, PRL 124 (2020) 182001

Unfold by iterating: OmniFold



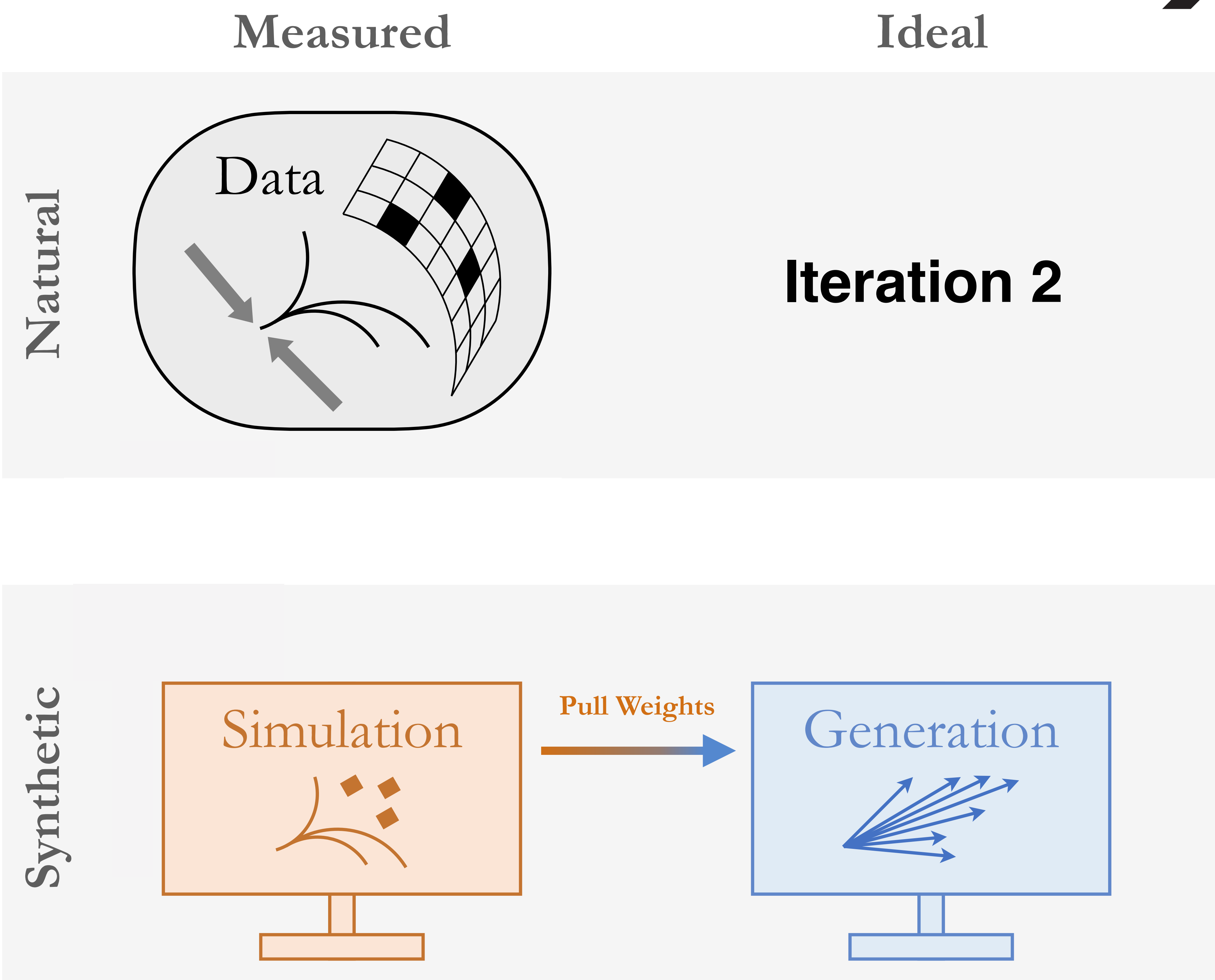
A. Andreassen, P. Komiske, E. Metodiev, BPN, J. Thaler, PRL 124 (2020) 182001

Unfold by iterating: OmniFold



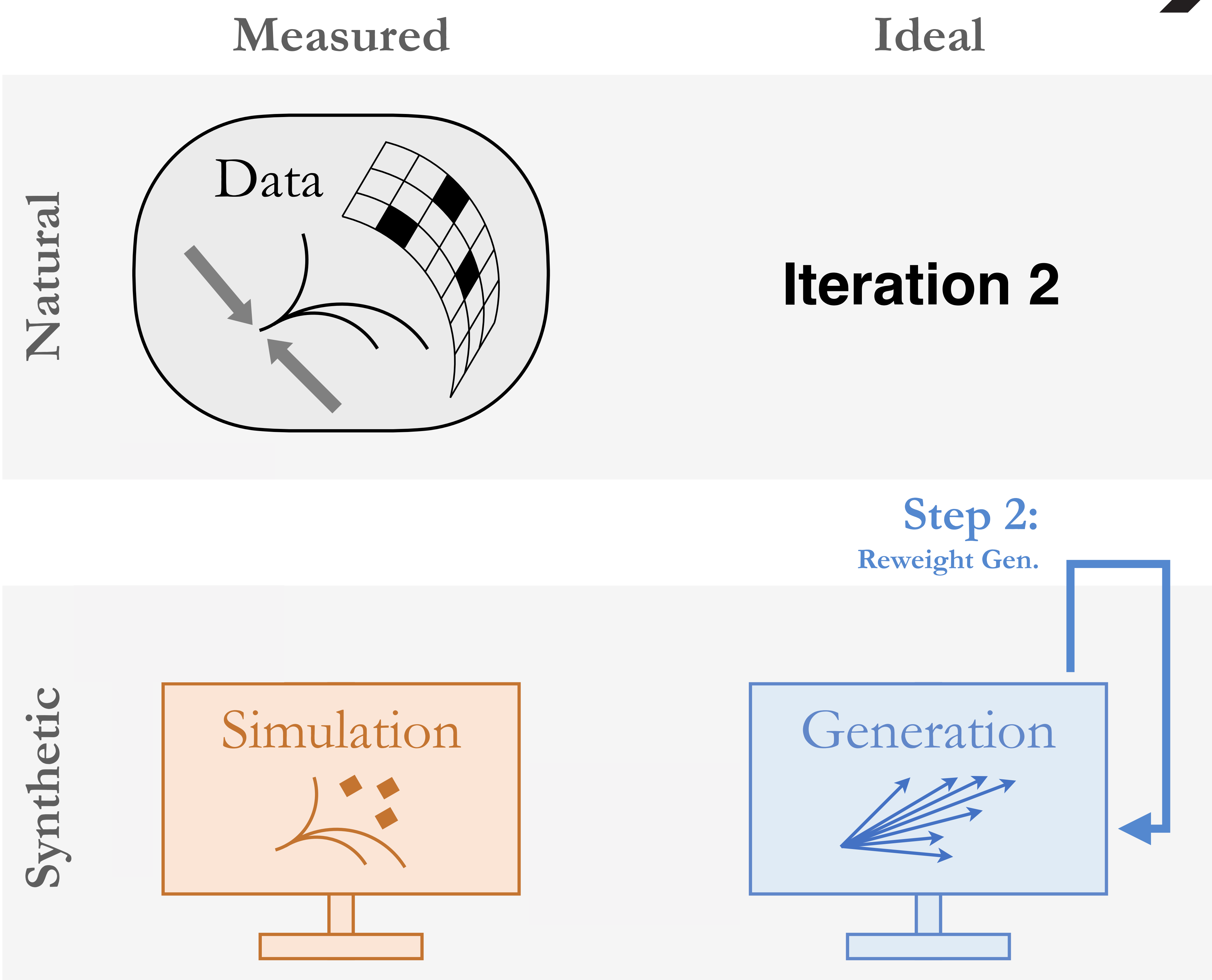
A. Andreassen, P. Komiske, E. Metodiev, BPN, J. Thaler, PRL 124 (2020) 182001

Unfold by iterating: OmniFold



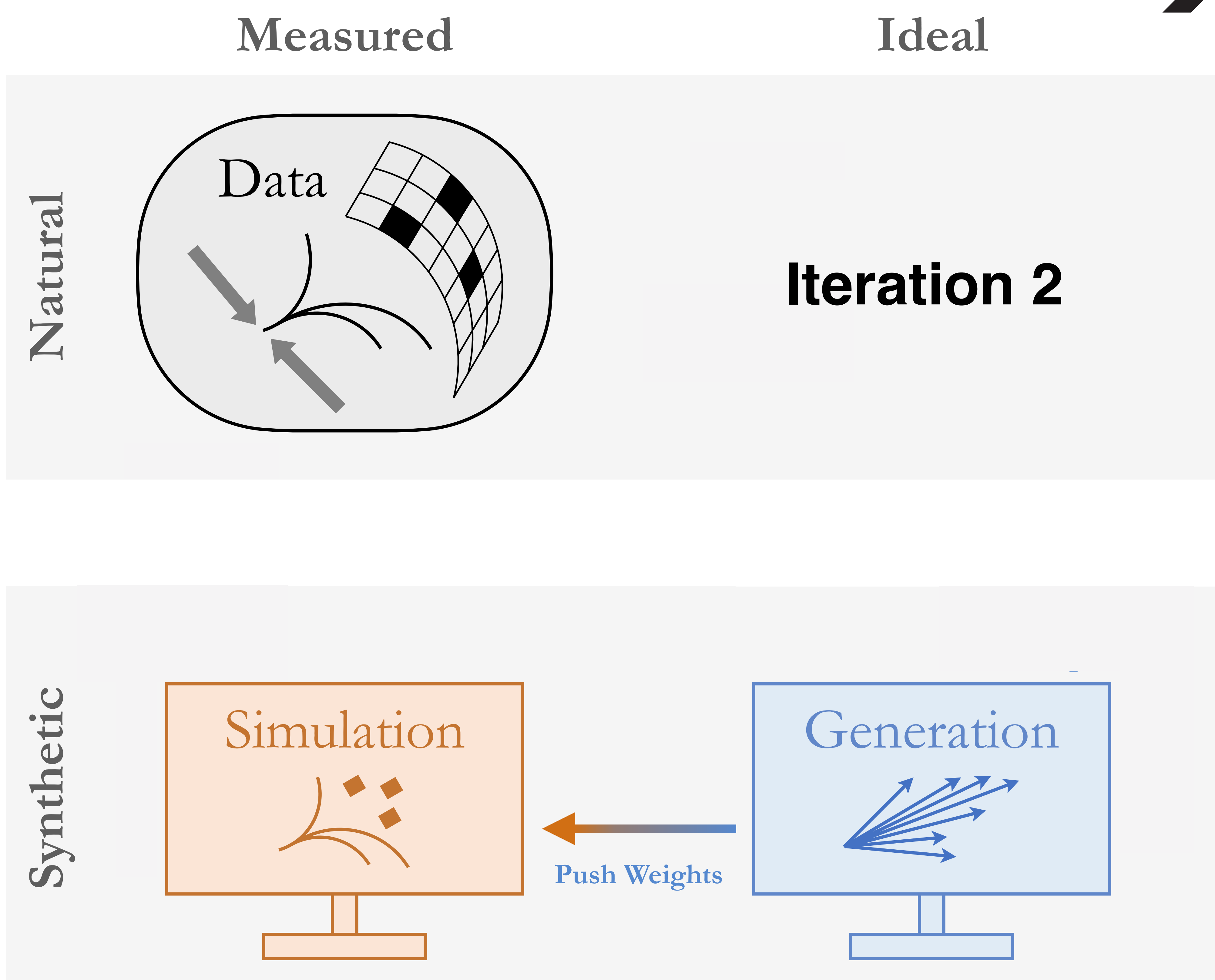
A. Andreassen, P. Komiske, E. Metodiev, BPN, J. Thaler, PRL 124 (2020) 182001

Unfold by iterating: OmniFold



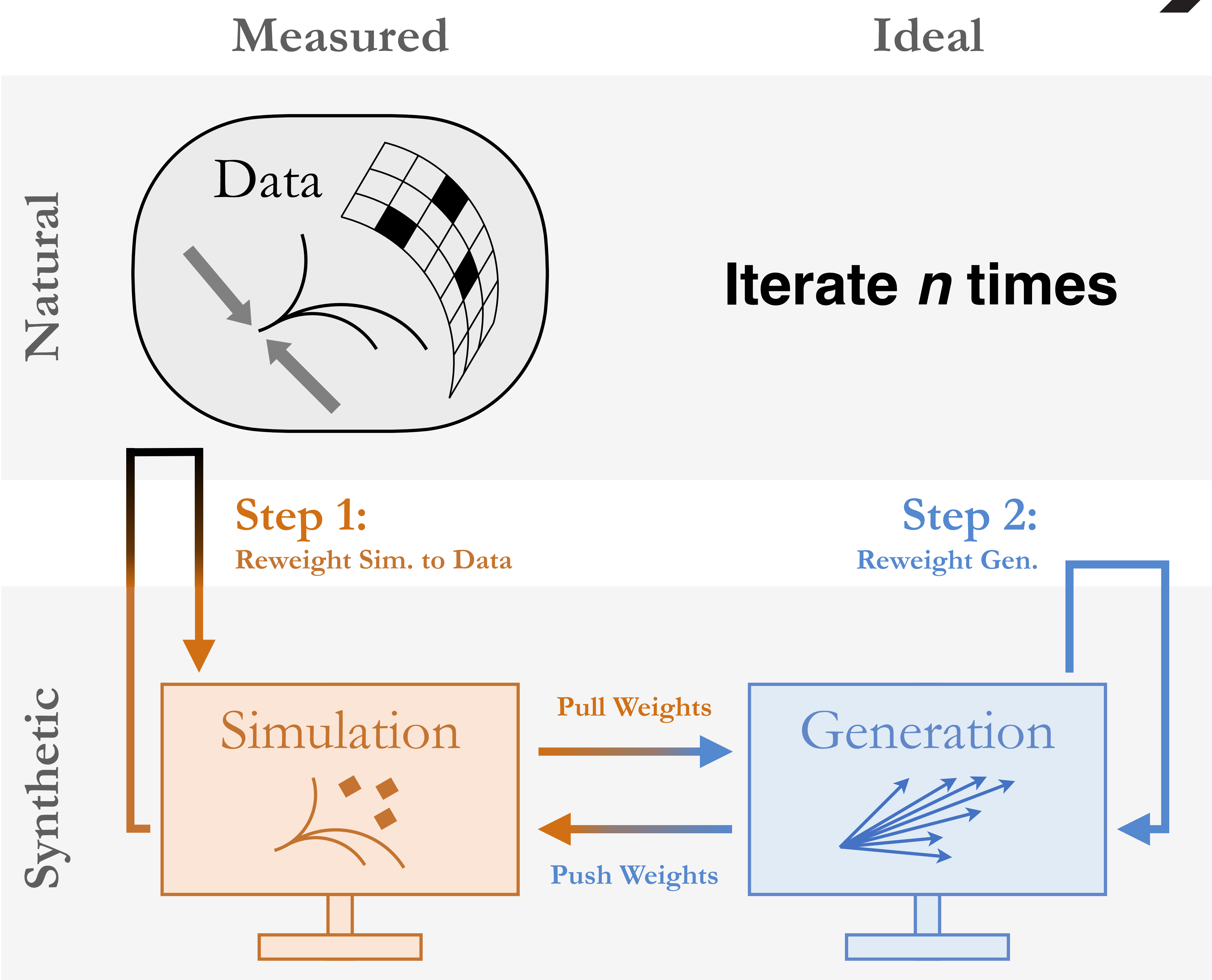
A. Andreassen, P. Komiske, E. Metodiev, BPN, J. Thaler, PRL 124 (2020) 182001

Unfold by iterating: OmniFold



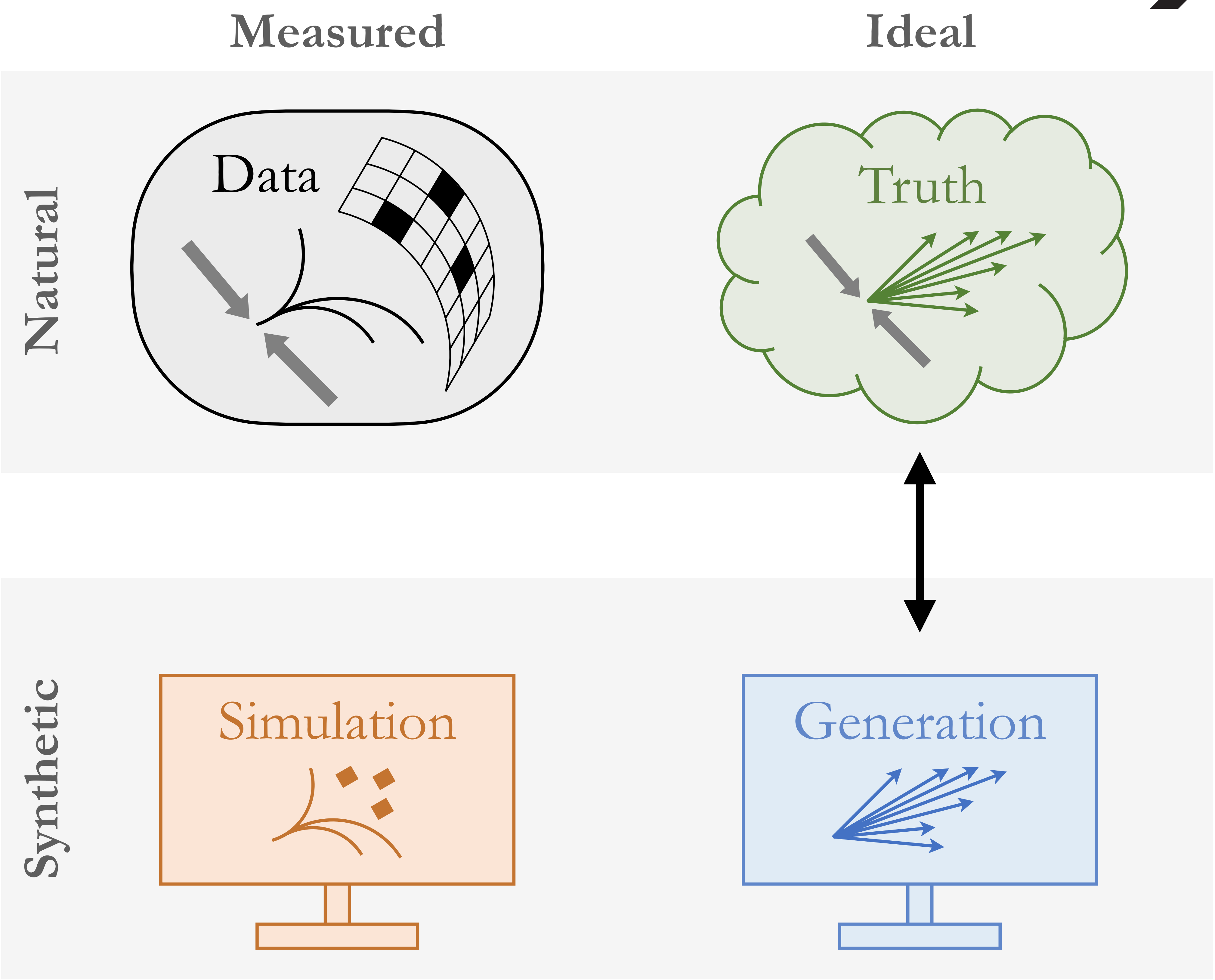
A. Andreassen, P. Komiske, E. Metodiev, BPN, J. Thaler, PRL 124 (2020) 182001

Unfold by iterating: OmniFold



A. Andreassen, P. Komiske, E. Metodiev, BPN, J. Thaler, PRL 124 (2020) 182001

Unfold by iterating: OmniFold



A. Andreassen, P. Komiske, E. Metodiev, BPN, J. Thaler, PRL 124 (2020) 182001

Reweighting

How do to the reweighting without binning?

Reweighting

How do to the reweighting without binning?

dataset 1: sampled from ***p(x)***
dataset 2: sampled from ***q(x)***

Create weights ***w(x) = q(x)/p(x)*** so that when dataset 1 is weighted by ***w***, it is statistically identical to dataset 2.

What if we don't (and can't easily) know ***q*** and ***p***?

(and don't want to estimate them by binning)

Classification for reweighting

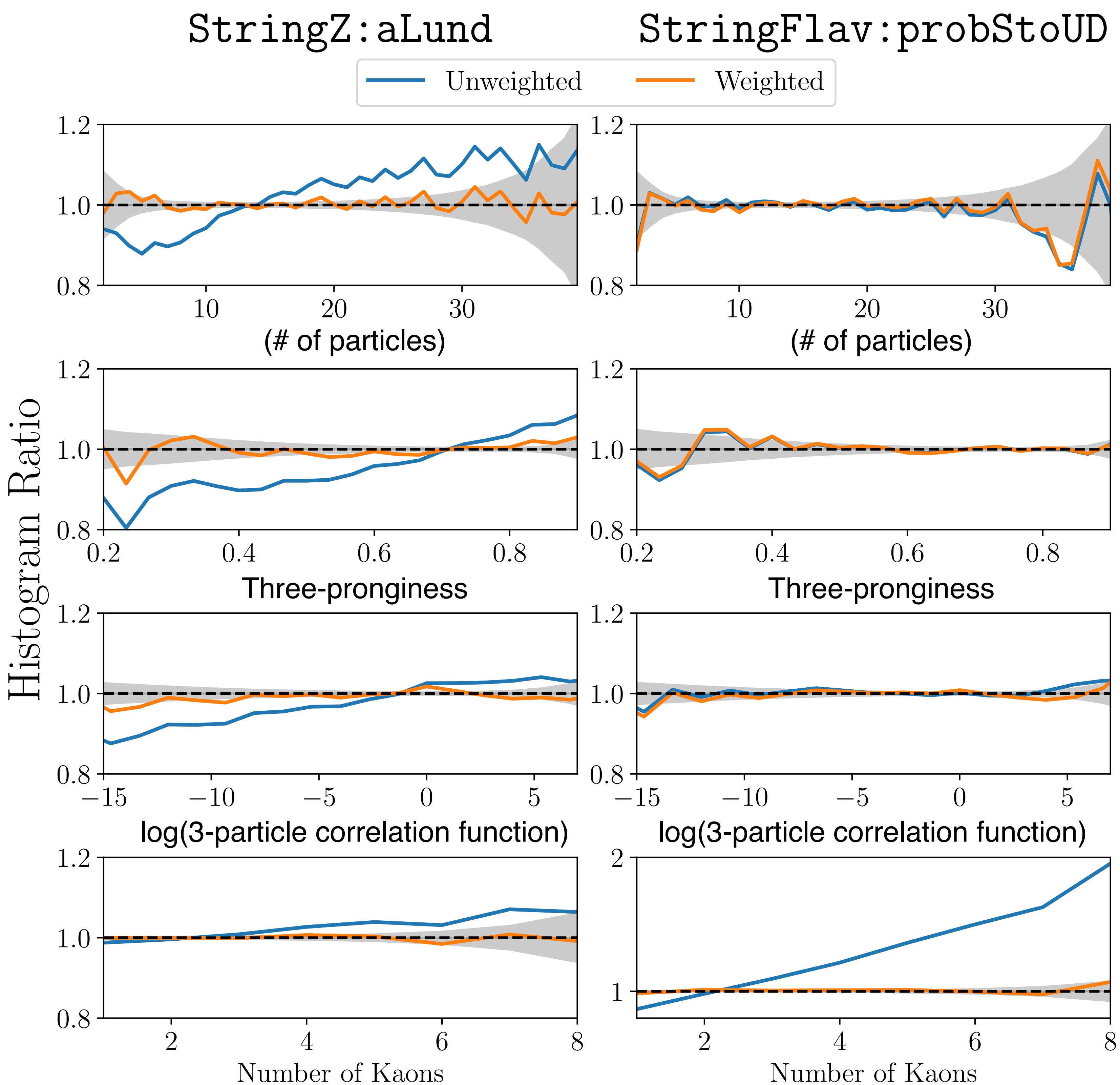
Fact: Neutral networks learn to approximate the likelihood ratio = $q(x)/p(x)$

(or something monotonically related to it in a known way)

Solution: train a neural network to distinguish the two datasets!

This turns the problem of **density estimation** (**hard**) into a problem of **classification** (**easy**)

Neural reweighing: works very well!



Full phase-space reweighing using simulated e^+e^-

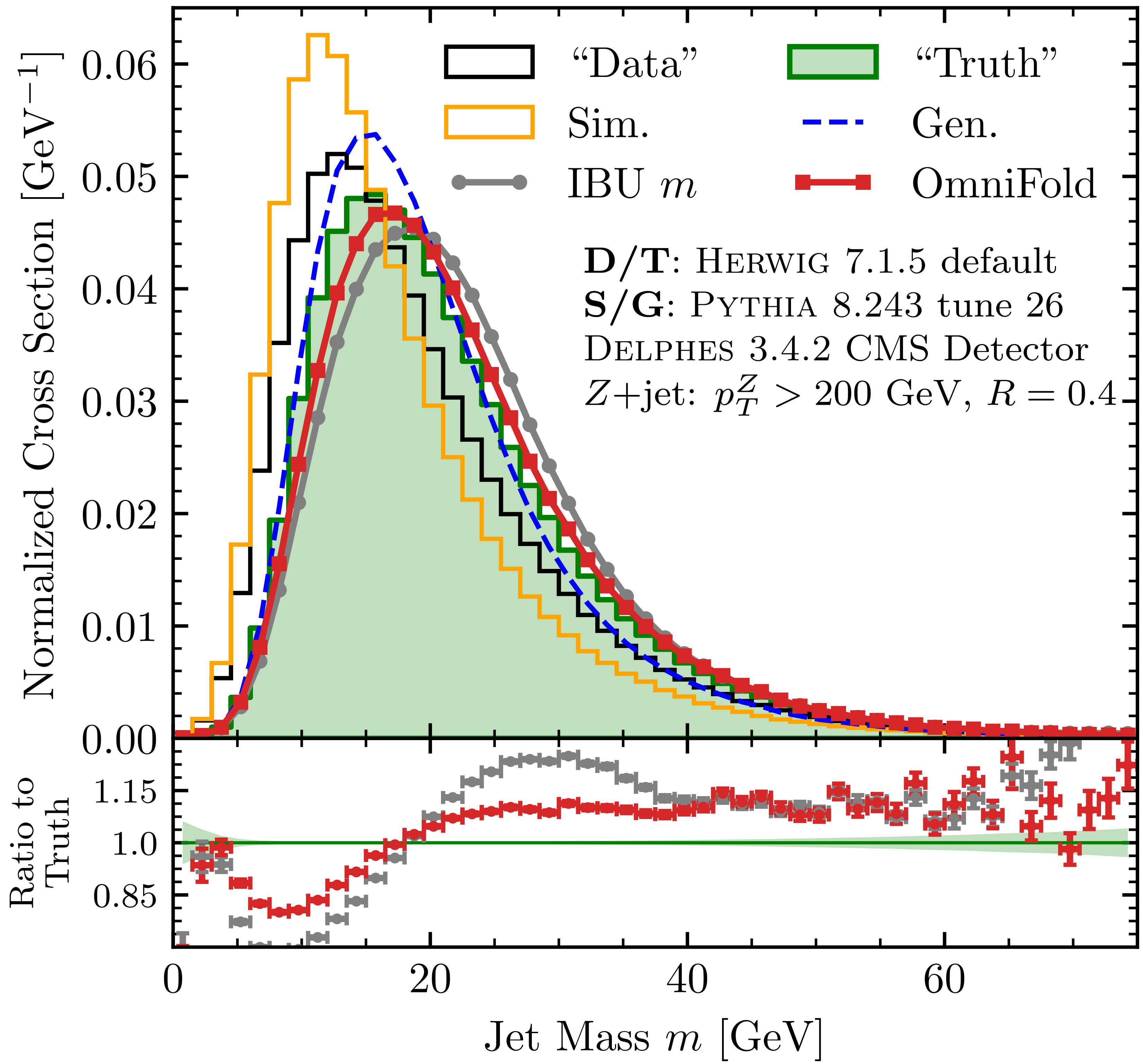
Works even when the differences are **small** (left) or **localized** (right).

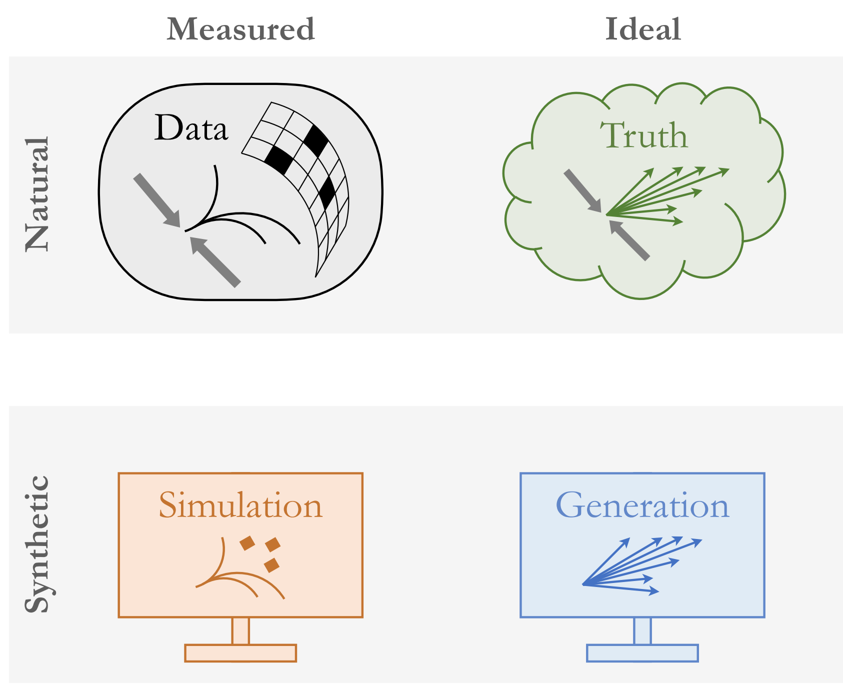
These are histogram ratios for a series of one-dimensional observables

A. Andreassen, BPN, PRD 101 (2020) 091901

Full phase-space unfolding

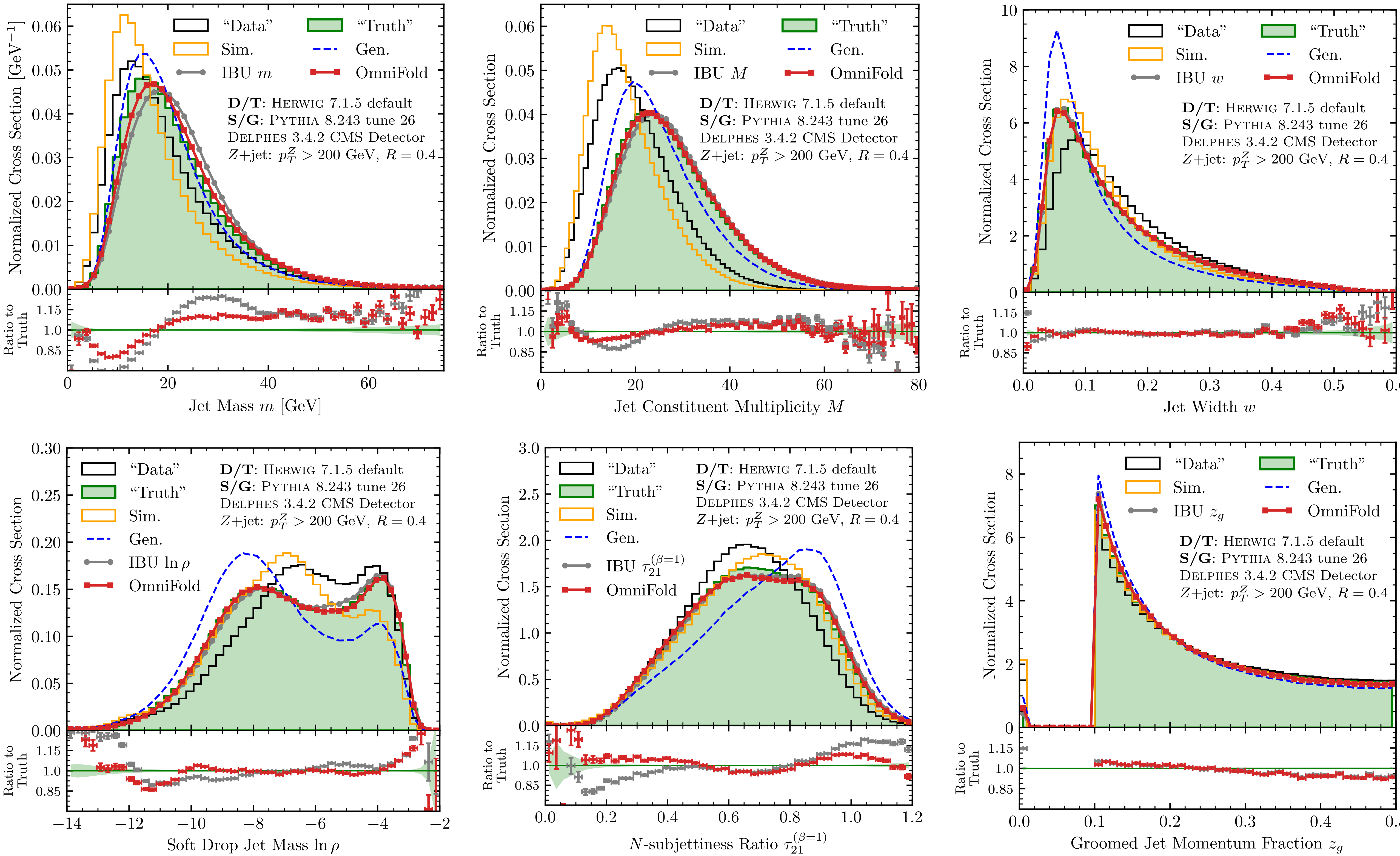
A. Andreassen, P. Komiske, E. Metodiev, BPN, J. Thaler, PRL 124 (2020) 182001





Full phase-space unfolding

A. Andreassen, P. Komiske, E. Metodiev, BPN, J. Thaler, PRL 124 (2020) 182001



Full phase-space unfolding

OmniFold is:

- *Unbinned*
- *Maximum likelihood**
- *Improves the resolution from correlations with detector response*

**when binned, OmniFold converges to Lucy-Richardson (aka Iterative Bayesian Unfolding)*

In fact, OmniFold can also work on low-level inputs (e.g. energy flow particles). In that case, you can construct observables ***after*** *the measurement.*

Some technical details

Please ask if you are interested, but briefly, OmniFold…

- Can accommodate backgrounds (unbinned) via neural positive reweighing
- Can accommodate acceptance effects
- Has a number of choices for how to update weights and/or keep track of acceptance effects

https://github.com/hep-lbdl/OmniFold

See A. Andreassen et al., ICLR SimDL for details [https://simdl.github.io/files/12.pdf]

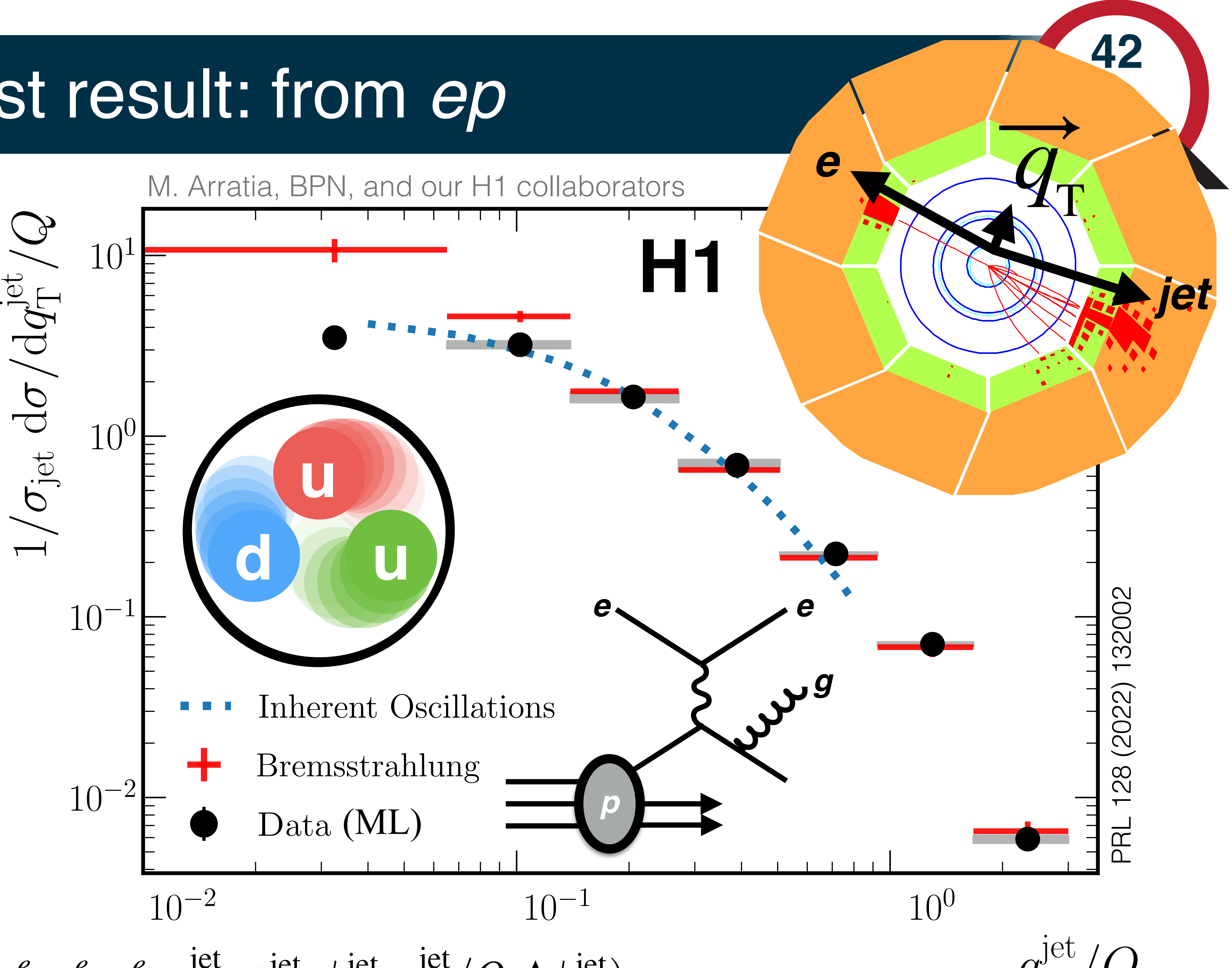
First Results

I'll now spend a few minutes flashing the first unbinned measurement results

There is no time to give the physics content justice, so I'll be brief, but please let me know if you have any questions!

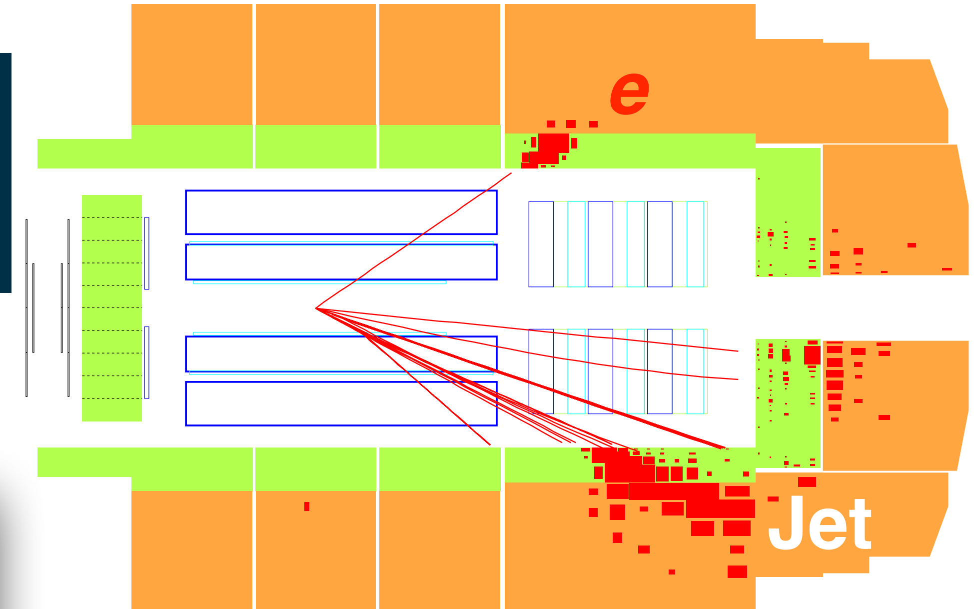
First result: from *ep*

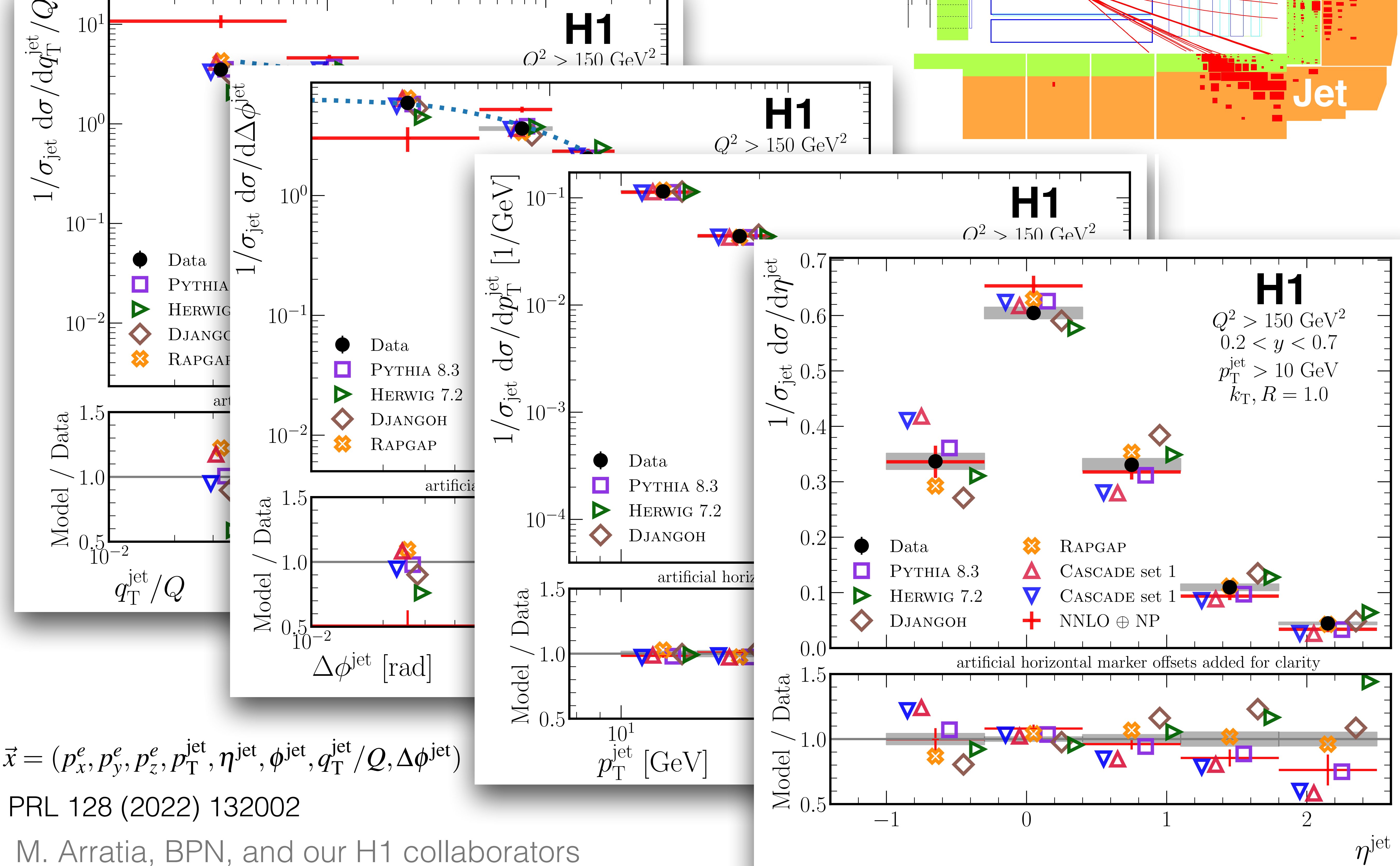
M. Arratia, BPN, and our H1 collaborators



$$\vec{x} = (p_x^e, p_y^e, p_z^e, p_T^{\rm jet}, \eta^{\rm jet}, \phi^{\rm jet}, q_T^{\rm jet}/Q, \Delta\phi^{\rm jet})$$

First result: from *ep*



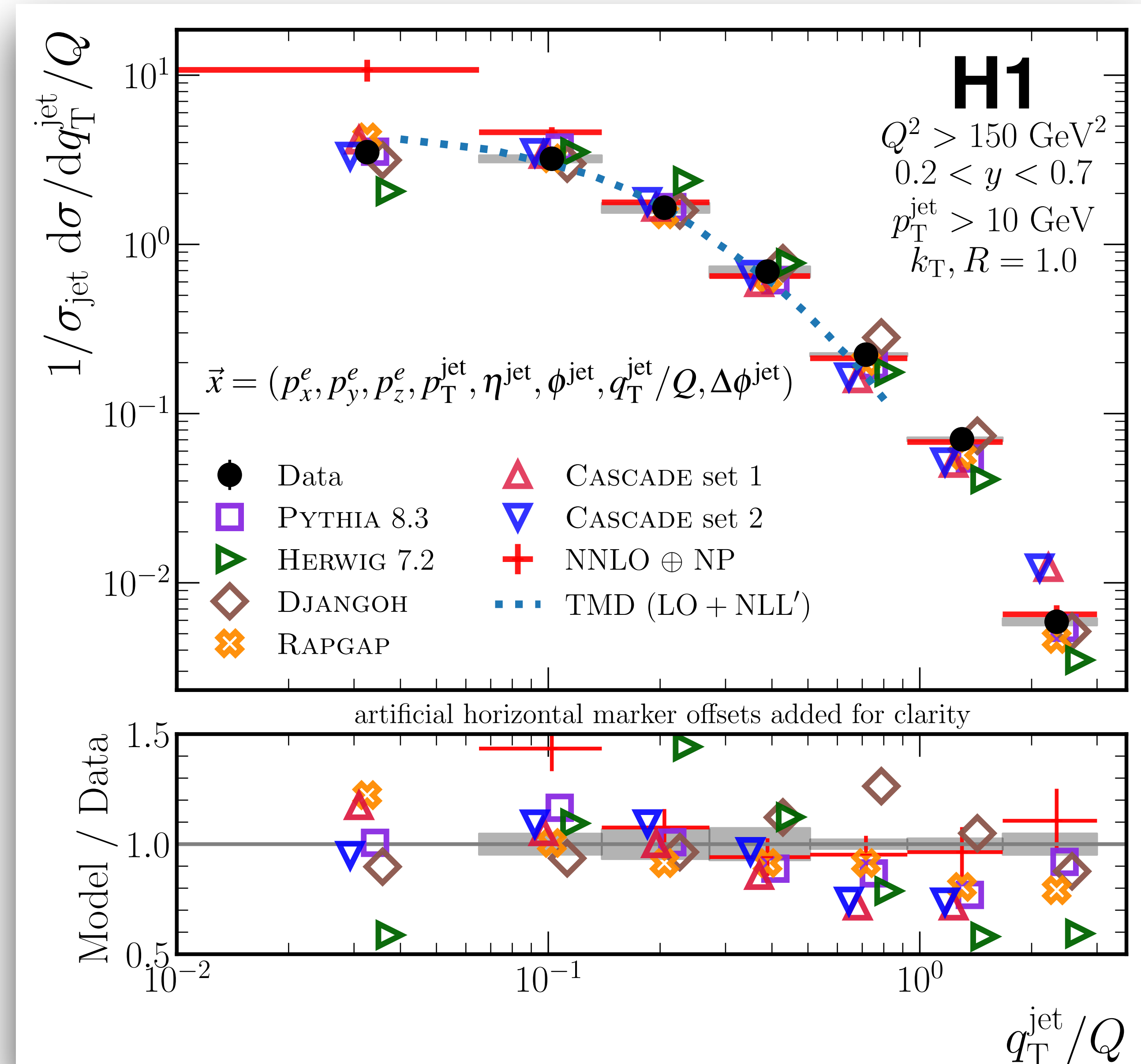


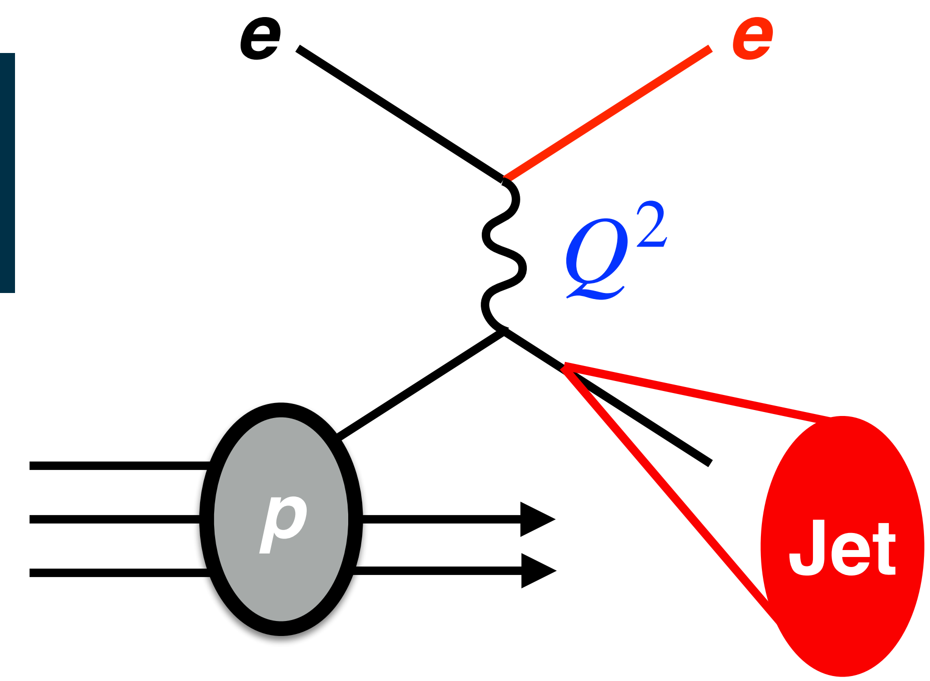
$$\vec{x} = (p_x^e, p_y^e, p_z^e, p_{\rm T}^{\rm jet}, \eta^{\rm jet}, \phi^{\rm jet}, q_{\rm T}^{\rm jet}/Q, \Delta\phi^{\rm jet})$$

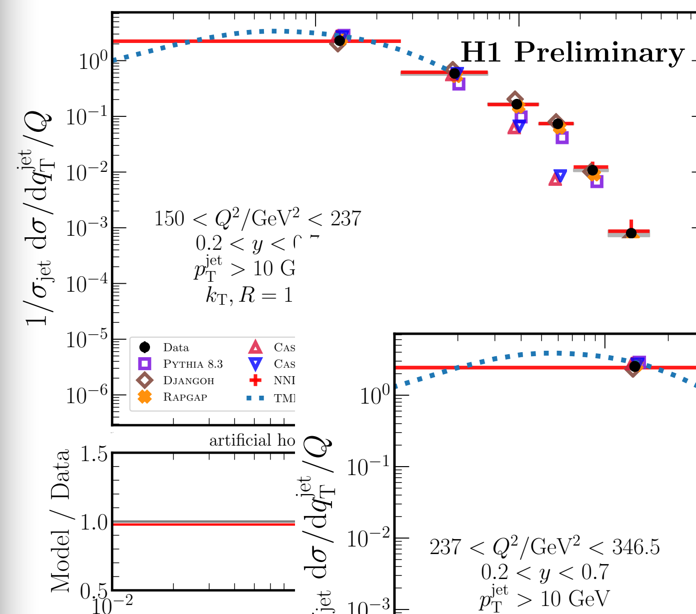
PRL 128 (2022) 132002

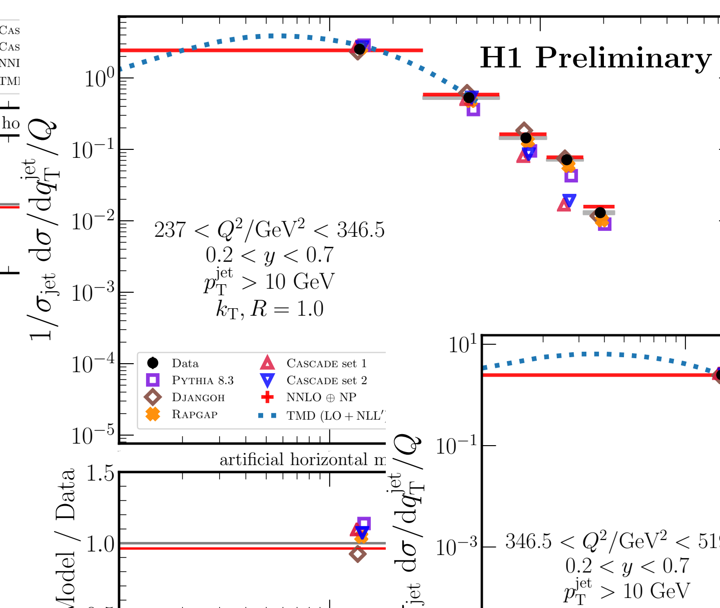
M. Arratia, BPN, and our H1 collaborators

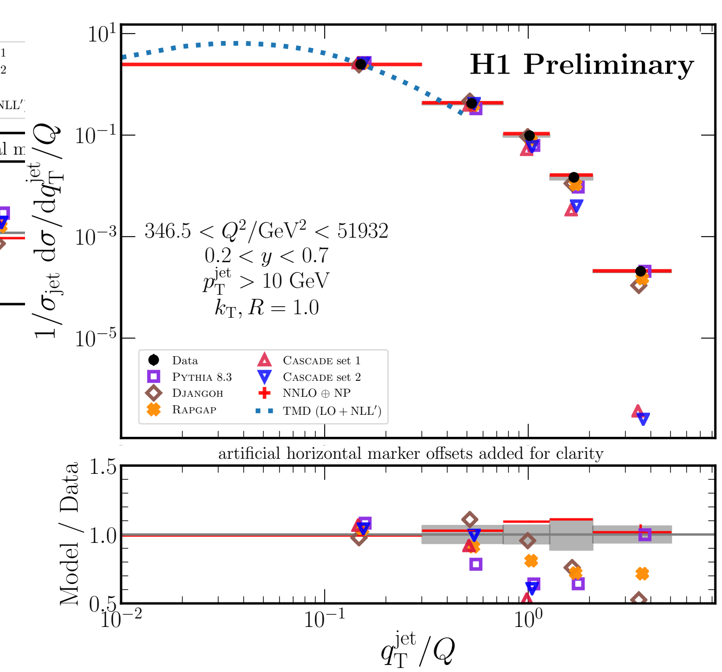
Re-using and extending











M. Arratia, BPN, Y. Xu and our H1 collaborators

H1prelim-22-031

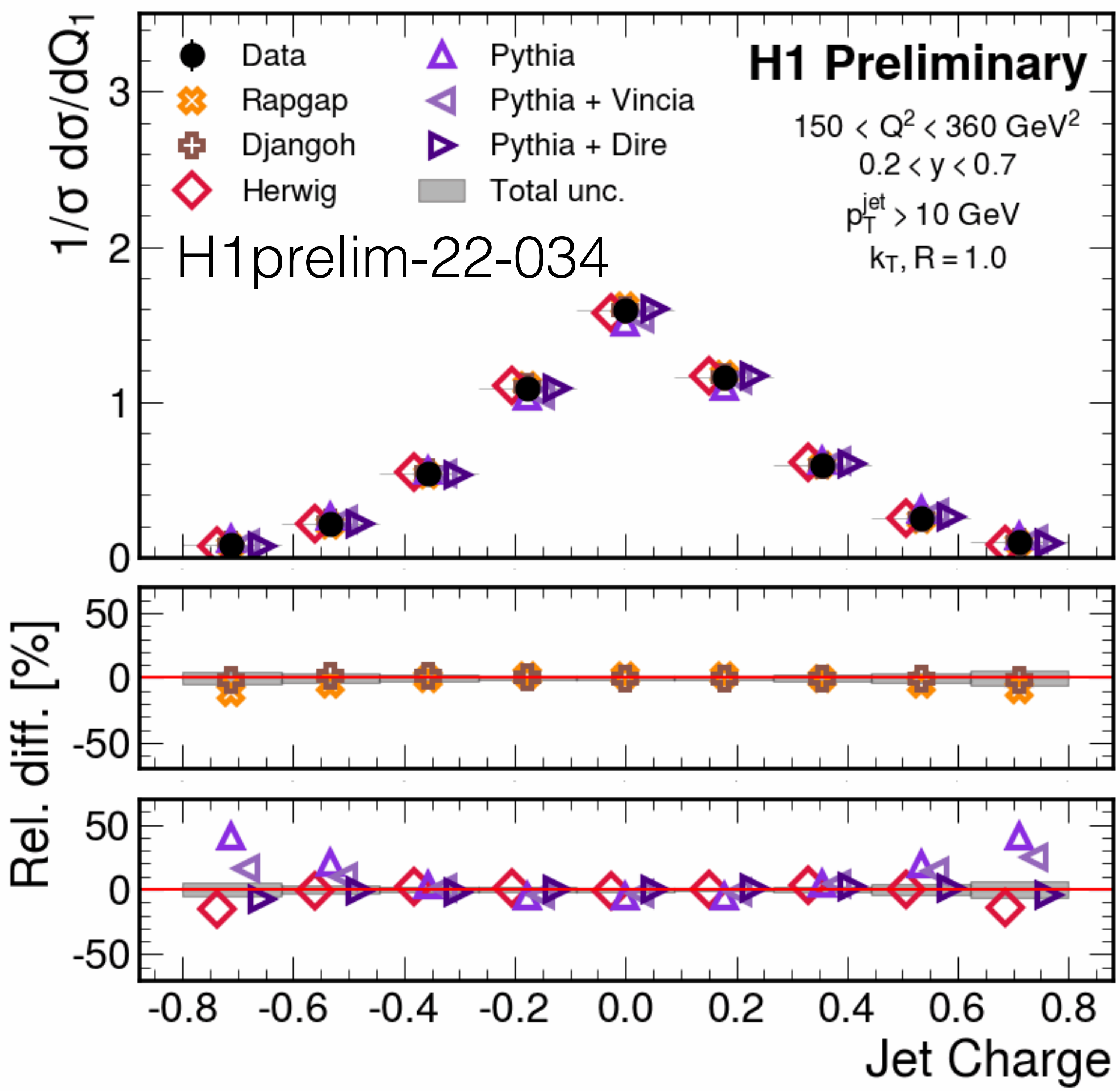
Increasing Q^2

$$Q^2 = \frac{(q_T^{jet})^2}{(q_T^{jet}/Q)^2}$$

$$= \frac{(p_x^e + p_T^{jet}\cos(\phi^{jet}))^2 + (p_y^e + p_T^{jet}\sin(\phi^{jet}))^2}{(q_T^{jet}/Q)^2}$$

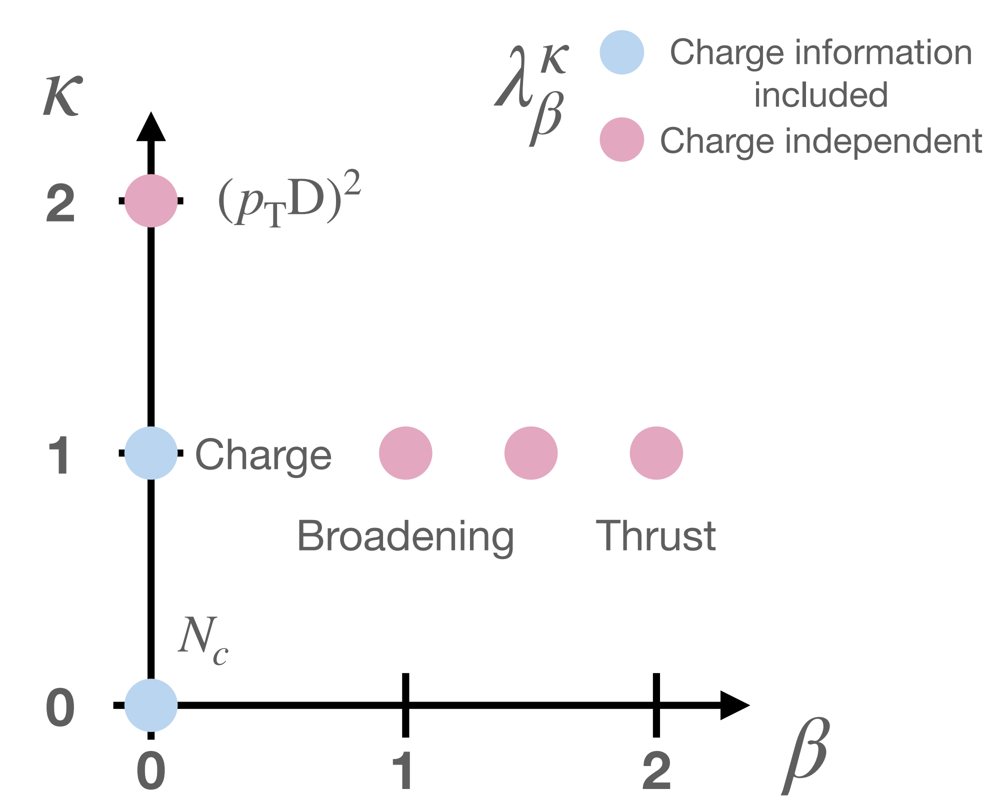
Looking inside jets

V. Mikuni, BPN, and our H1 collaborators



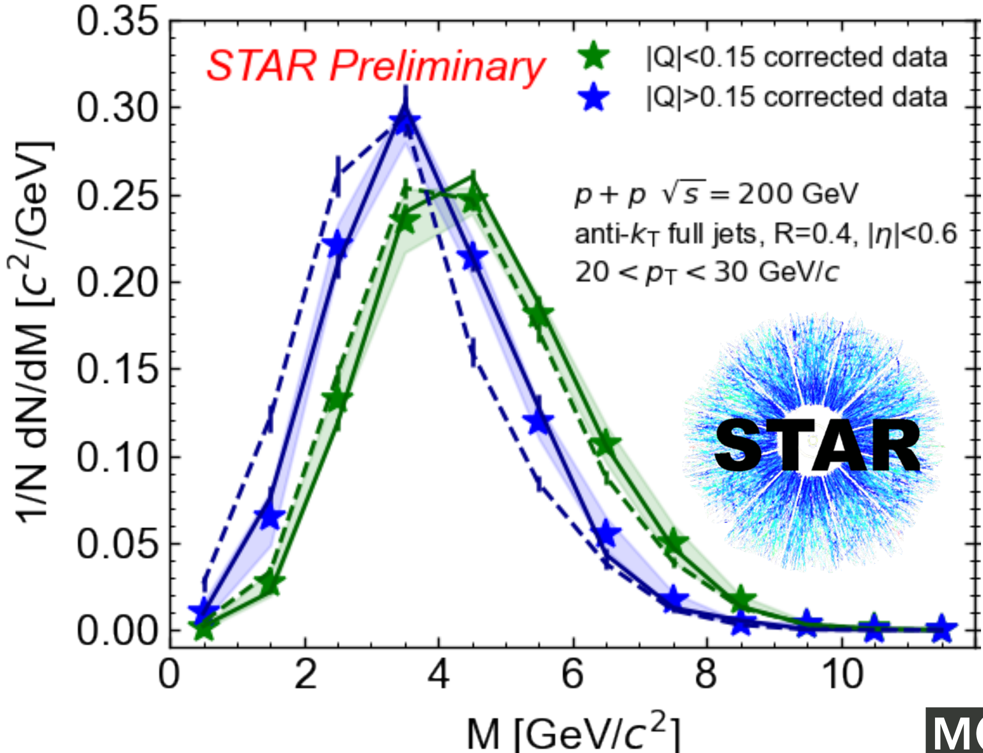
We have also studied high-dimensional data via graph neural networks*

(particle-level is low-D, but detector-level is high-D)

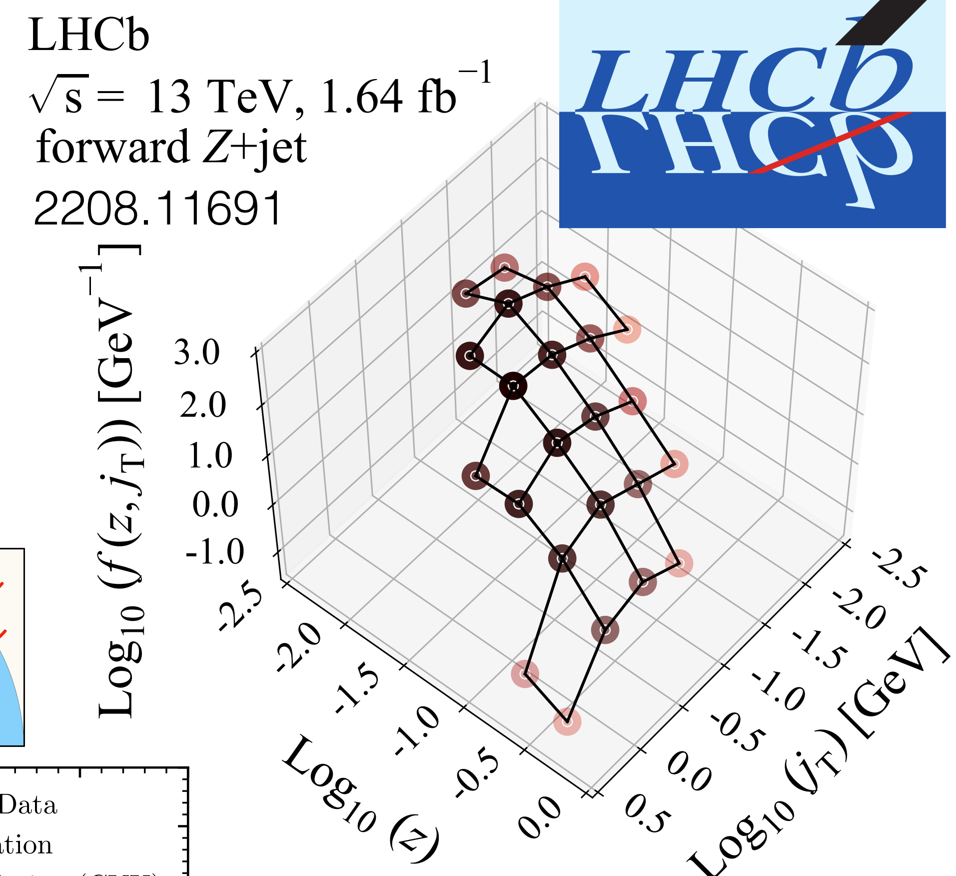


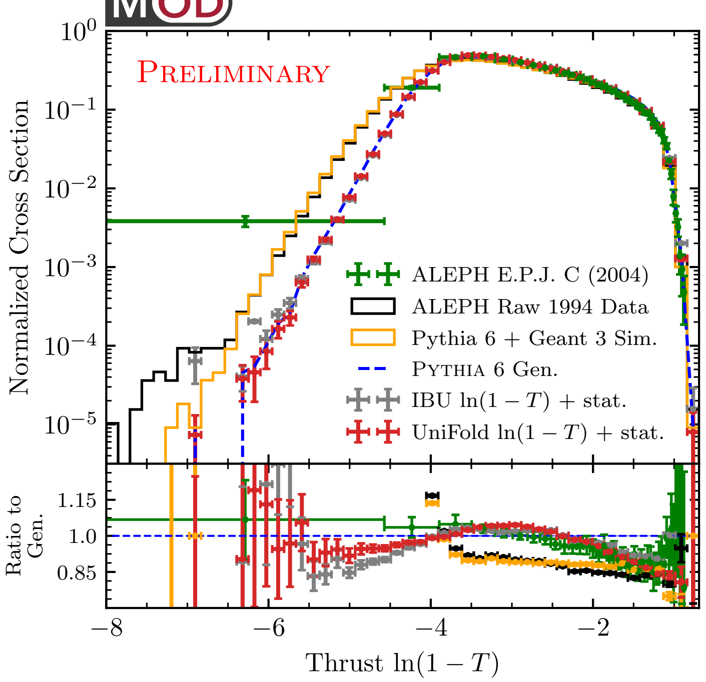
*M. Guo et al., CVM 7 (2021) 187; V. Mikuni, F. Canelli, MLST 2 (2021) 035027

Other studies + measurements w/data

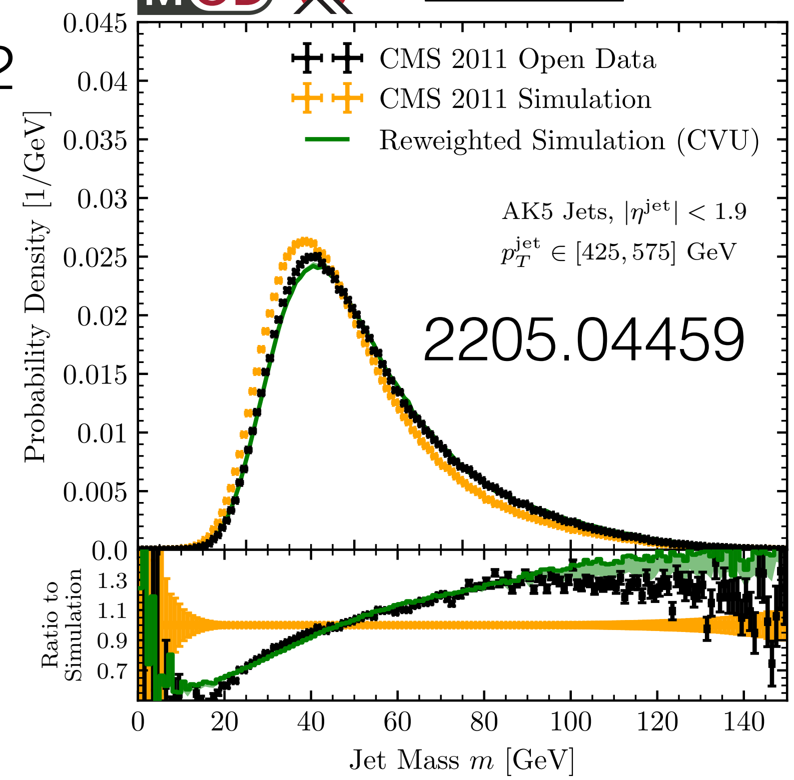


Y. Song (for STAR), DNP 2022

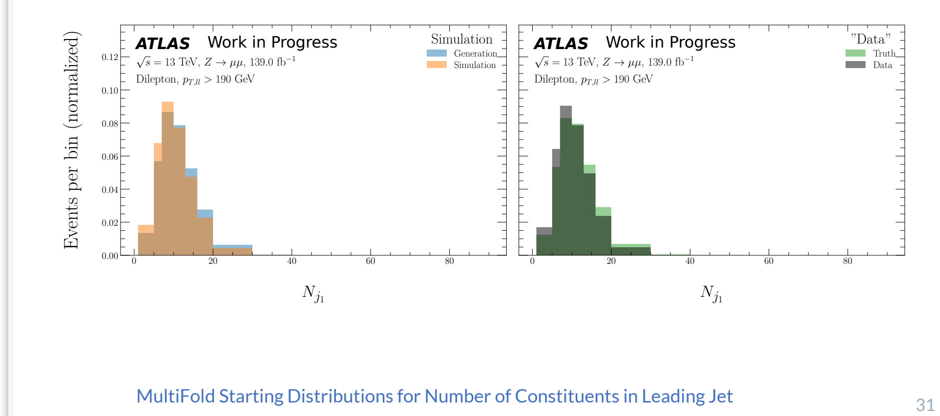




A. Badea et al., ICHEP 2020



A. Suresh (for ATLAS), DPF 2021



+others I know about, but don't have anything public yet + others I'm sure I don't know about!

Future + challenges

So far, OmniFold seems to work as designed!
Exciting to see where this will take us.

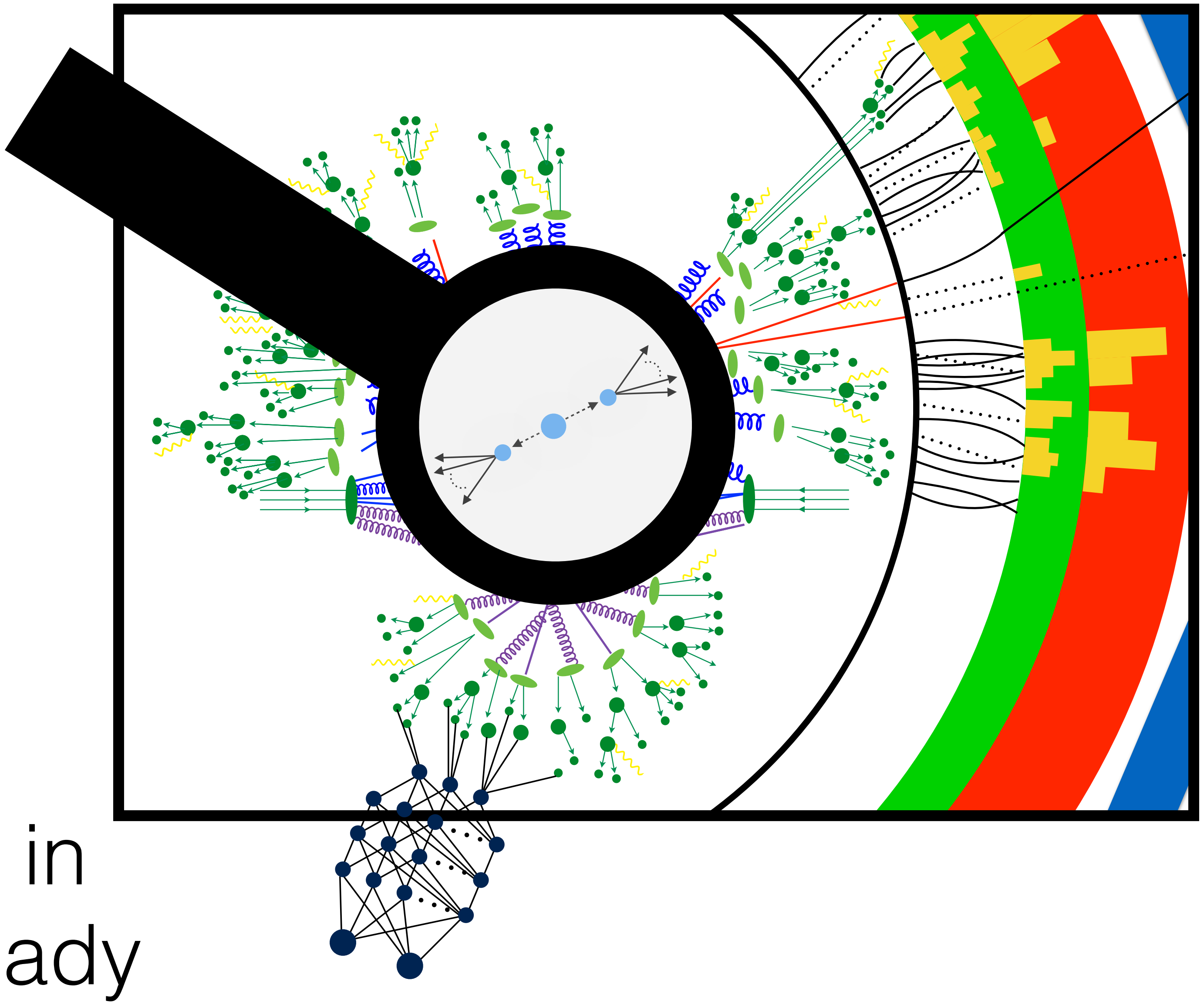
There are still some challenges we need to overcome:

- OmniFold is computationally expensive (need to train many networks, especially with ensembling to reach precision)
- How to publish an unbinned result? (all results so far are presented as binned) - see 2109.13243. Breaks HEPData!
- Modeling/closure uncertainties in high dimensions (not a new problem, but perhaps more acute)
- What about profiling? See 2302.05390 for a partial solution.

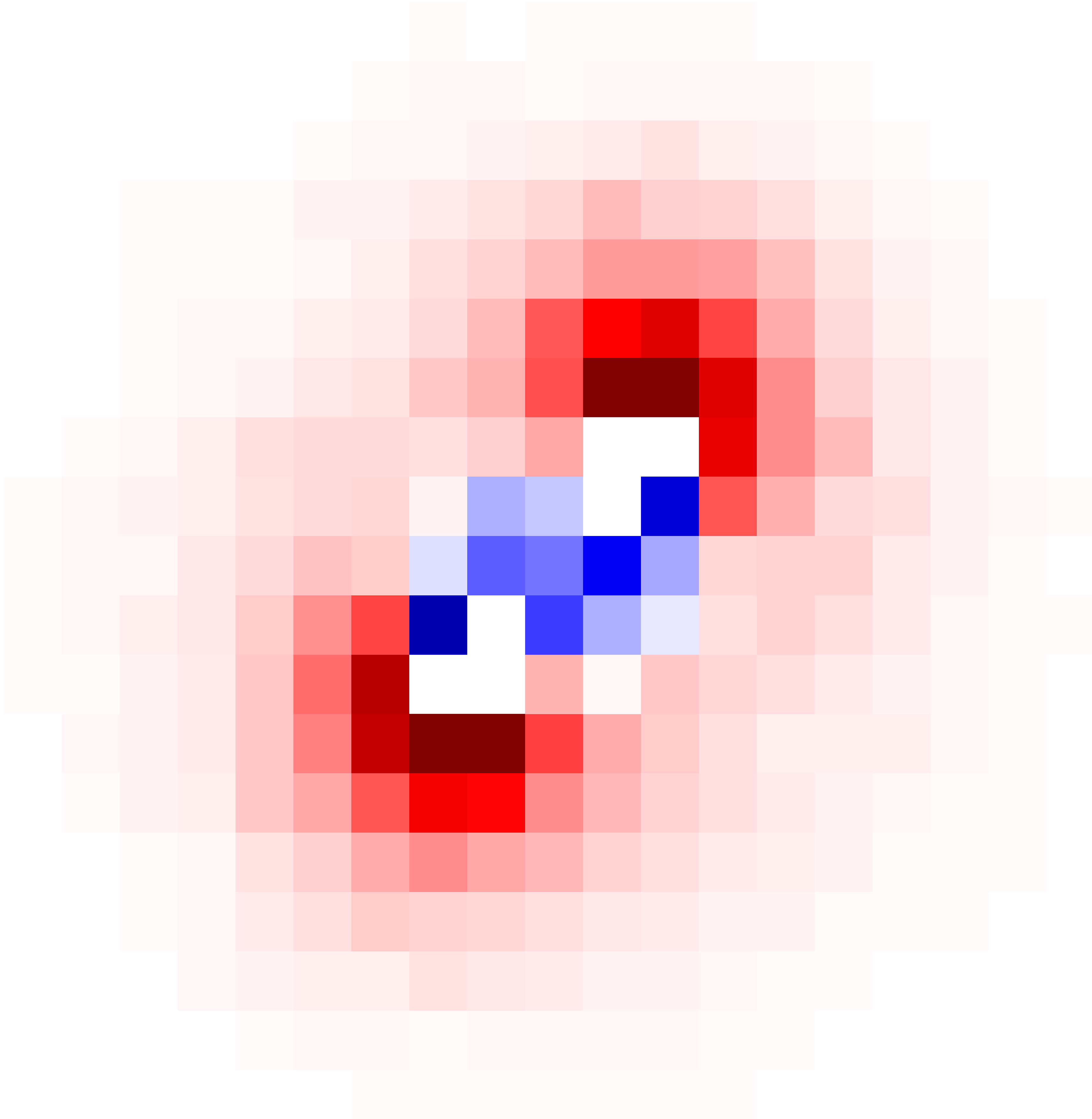
Conclusions and Outlook

A **new measurement paradigm** is possible, enabled by ML-based unfolding methods

We can analyze our data **holistically** and **future-proof** it using unbinned techniques

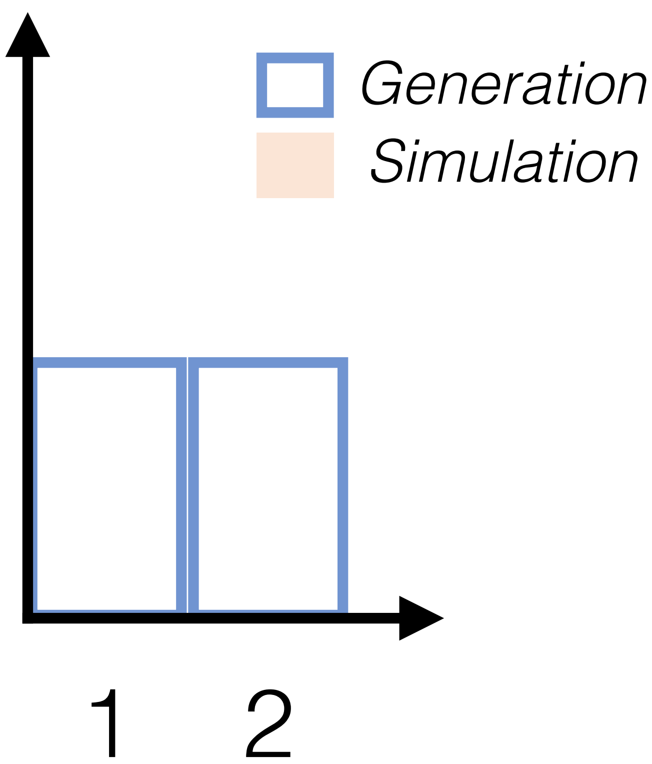


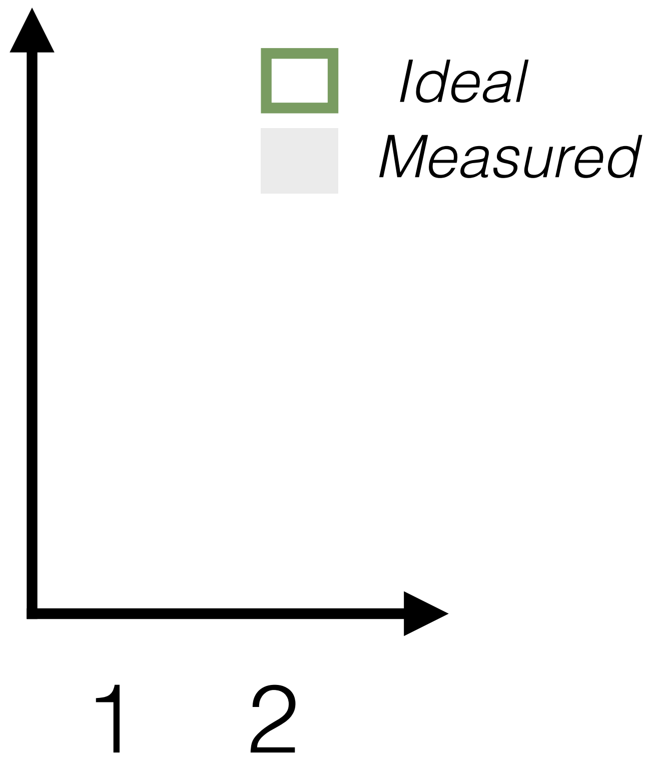
More R&D is required, but in parallel, these tools are already starting to **deliver science results**!

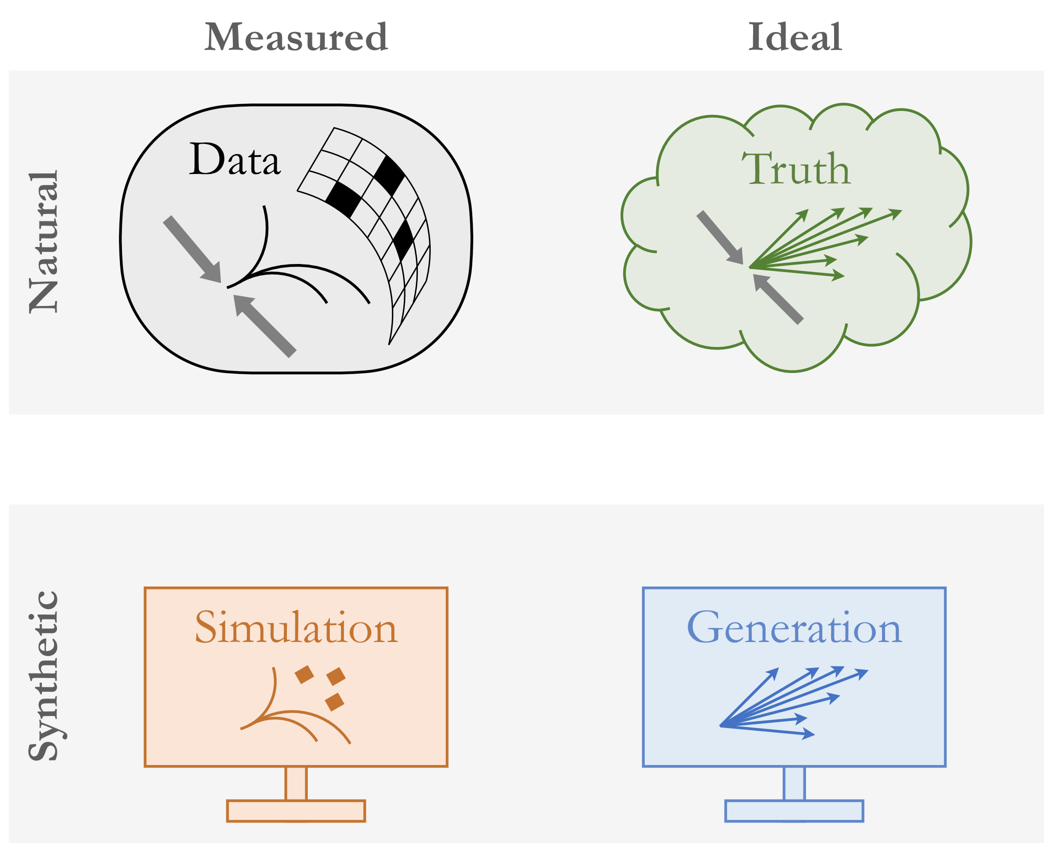


Fin.

Unfold by iterating: OmniFold

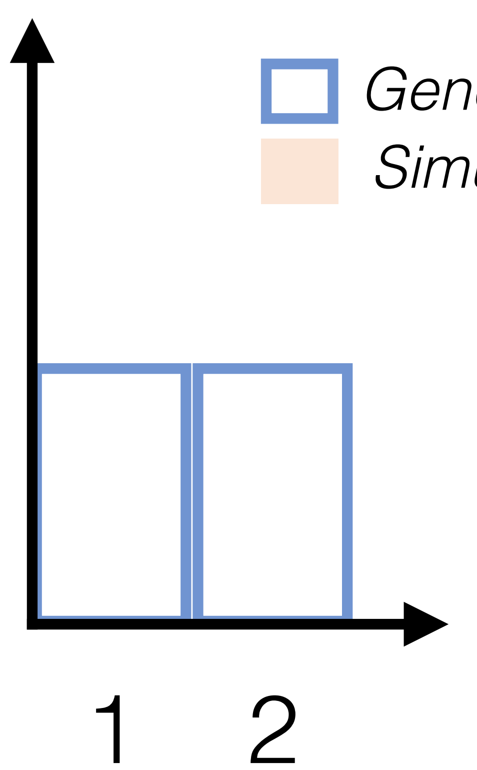


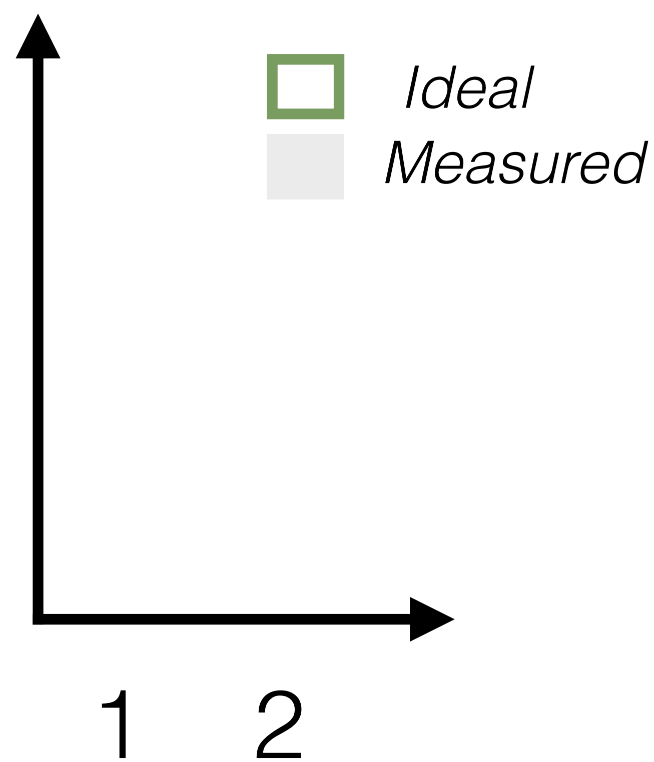


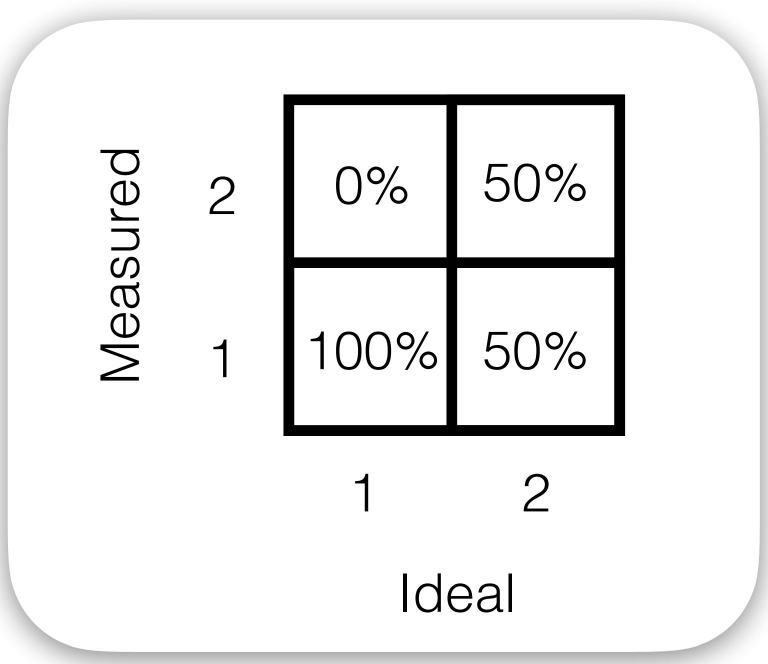


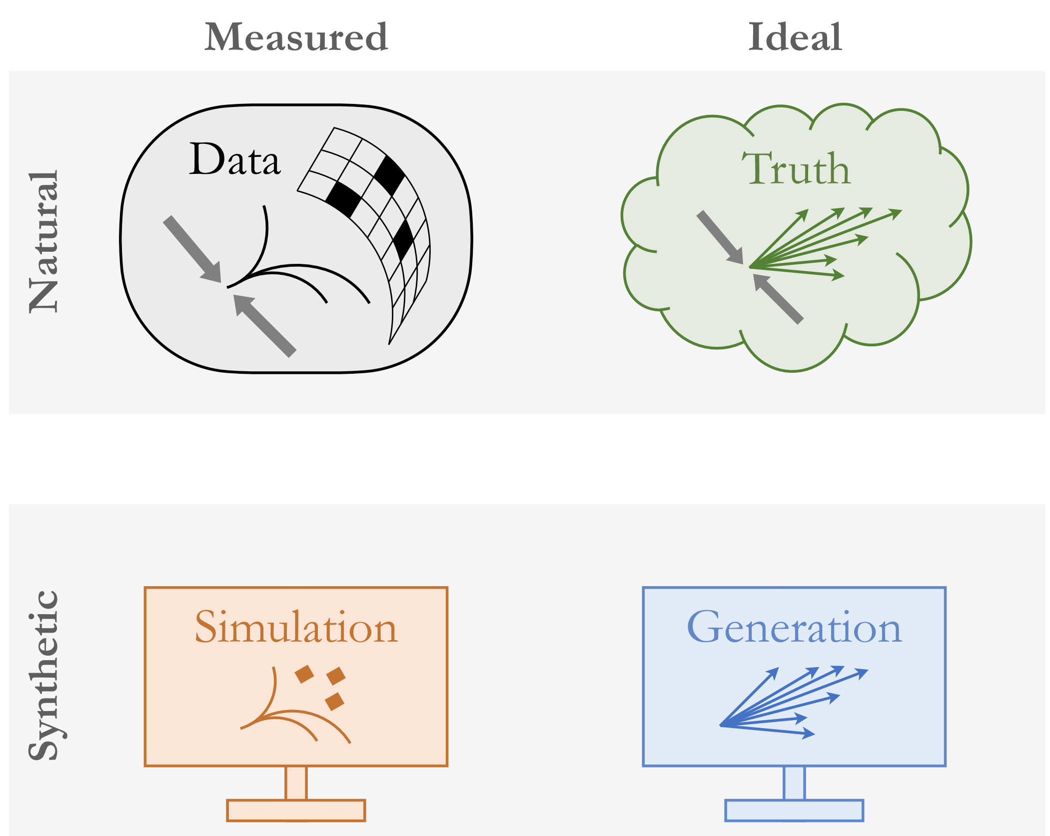
A. Andreassen, P. Komiske, E. Metodiev, BPN, J. Thaler, PRL 124 (2020) 182001

Unfold by iterating: OmniFold



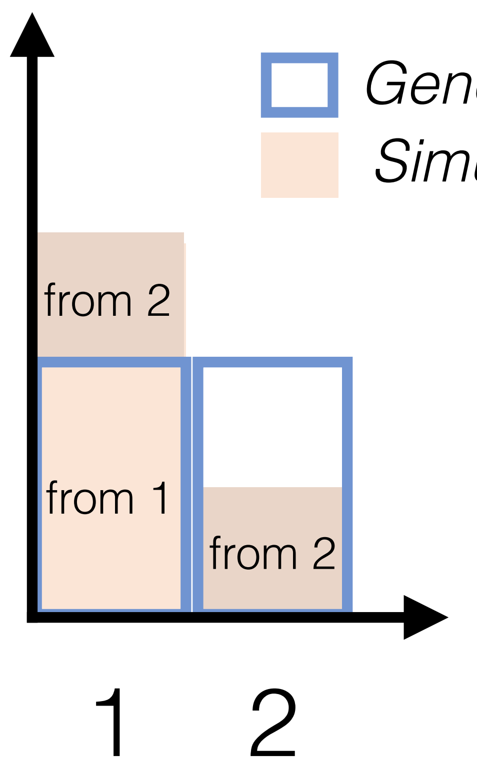


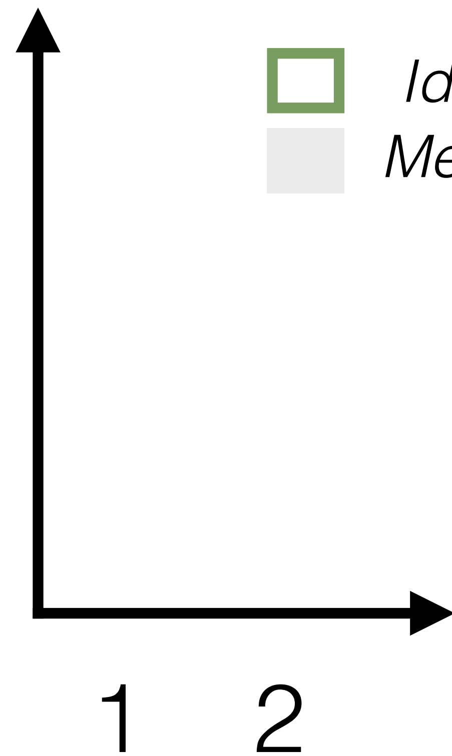




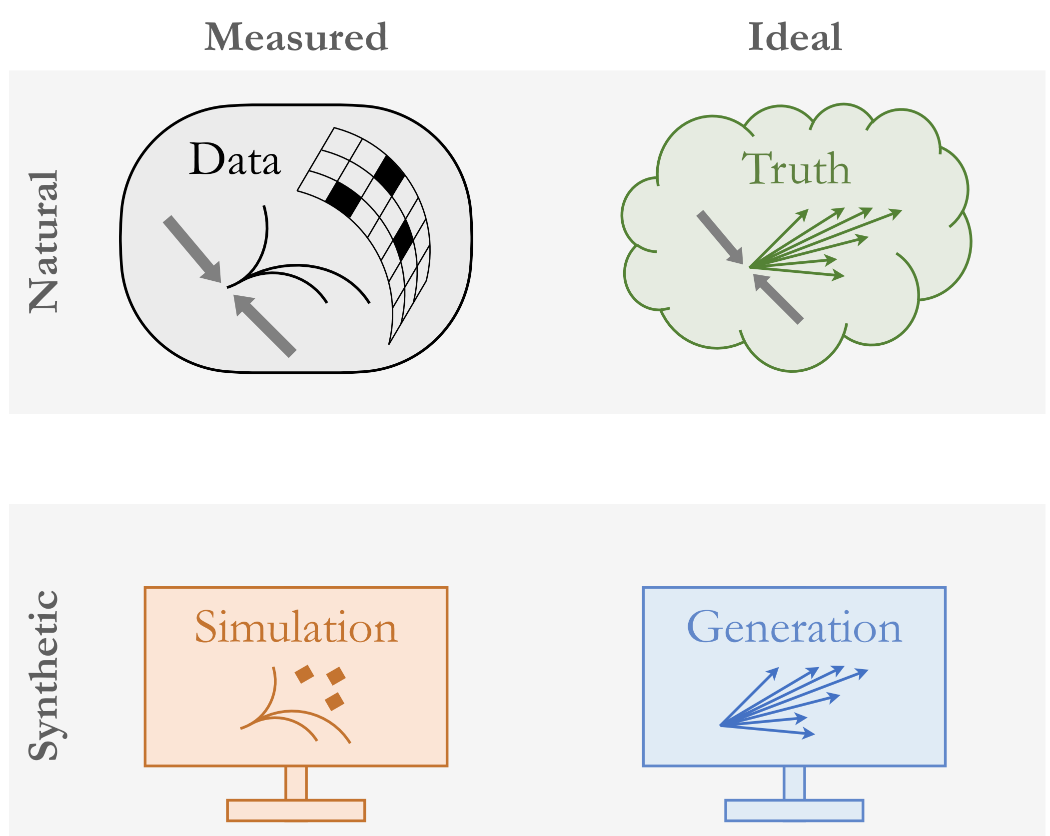
A. Andreassen, P. Komiske, E. Metodiev,
BPN, J. Thaler, PRL 124 (2020) 182001

Unfold by iterating: OmniFold



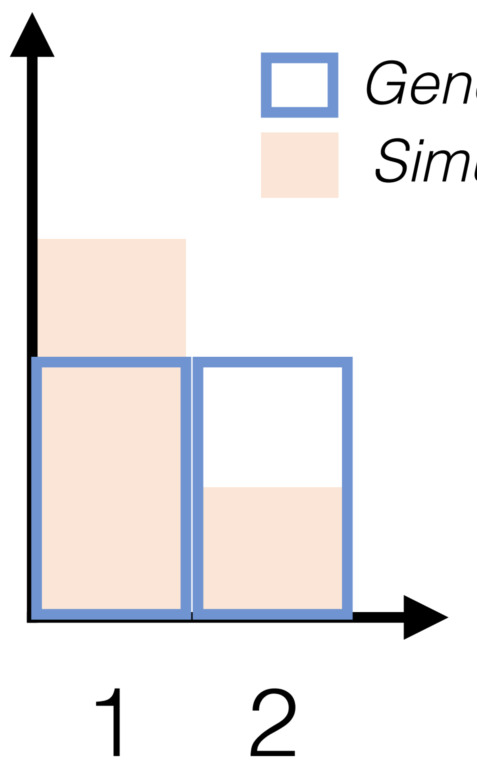


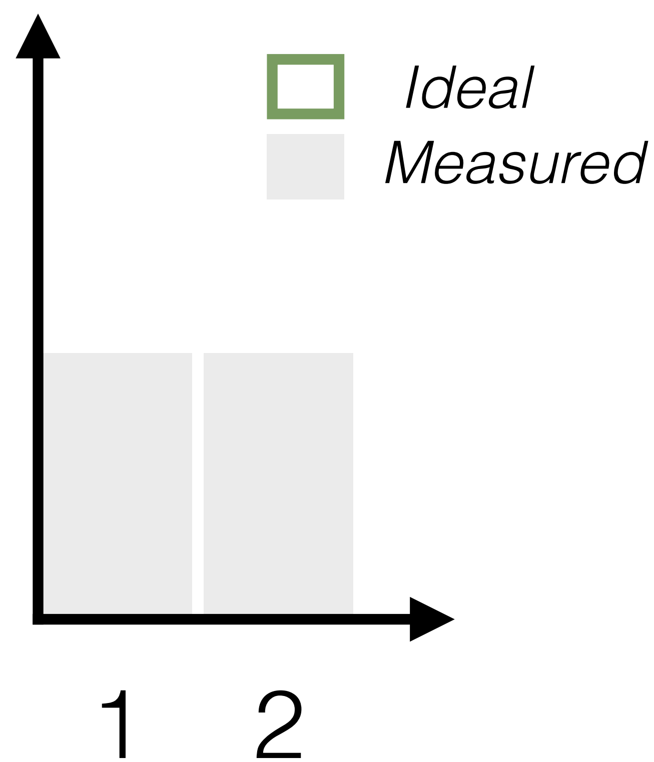
Measured \ Ideal	1	2
2	0%	50%
1	100%	50%



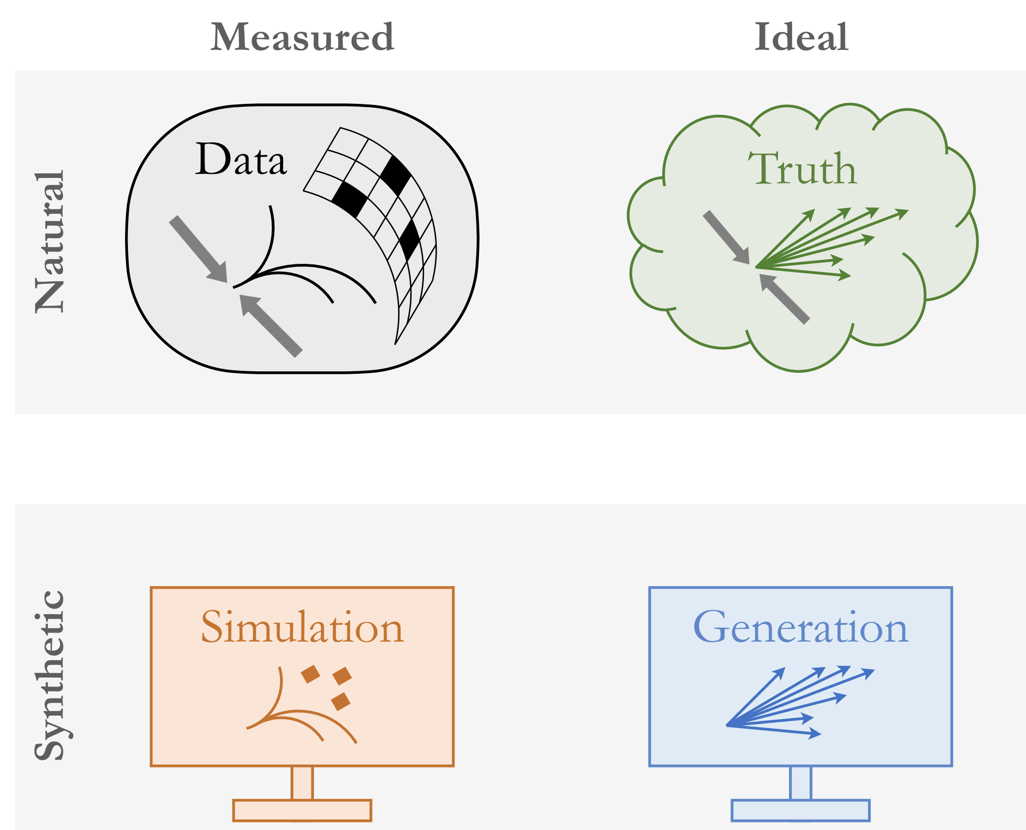
A. Andreassen, P. Komiske, E. Metodiev, BPN, J. Thaler, PRL 124 (2020) 182001

Unfold by iterating: OmniFold



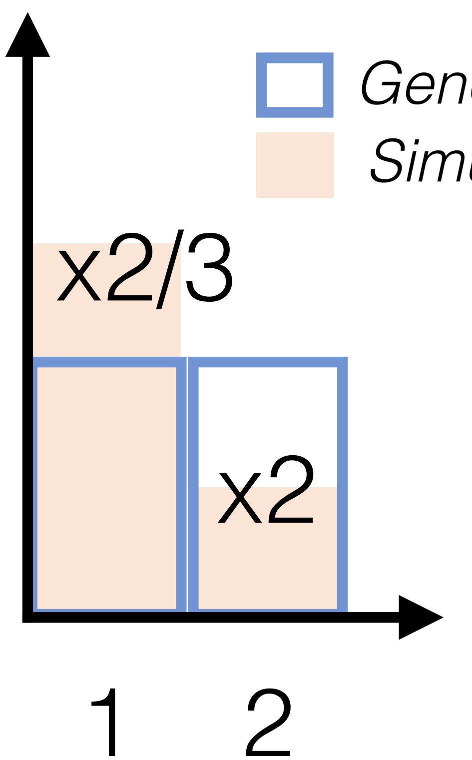


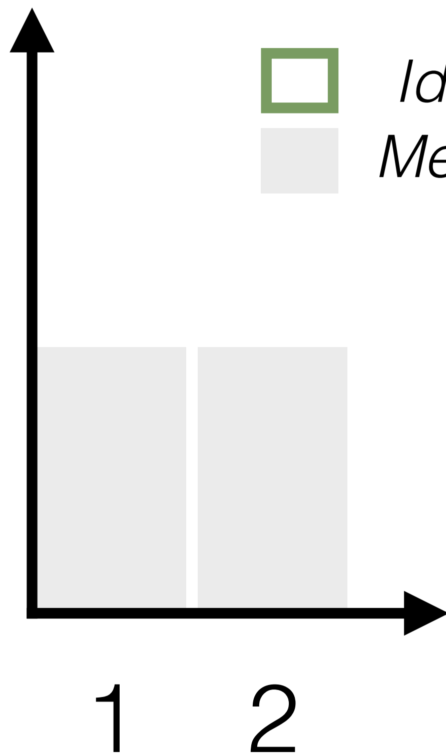
Measured \ Ideal	1	2
2	0%	50%
1	100%	50%

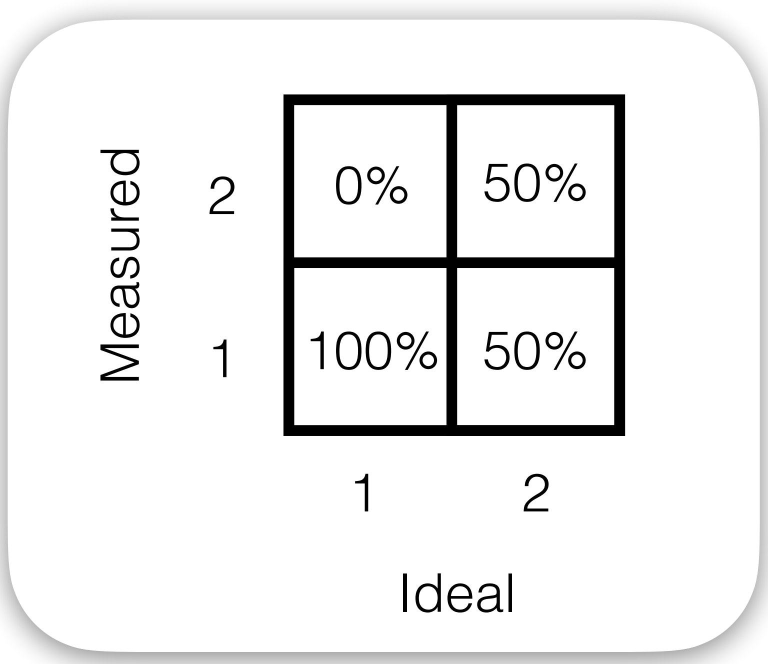


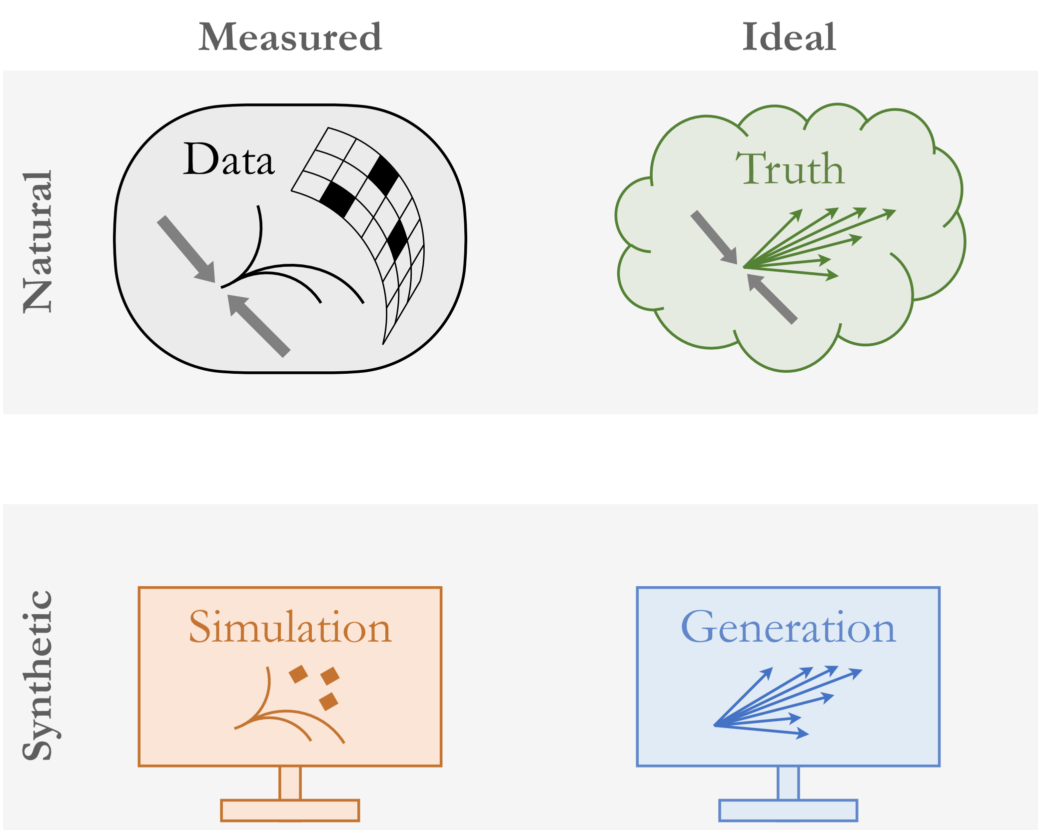
A. Andreassen, P. Komiske, E. Metodiev, BPN, J. Thaler, PRL 124 (2020) 182001

Unfold by iterating: OmniFold



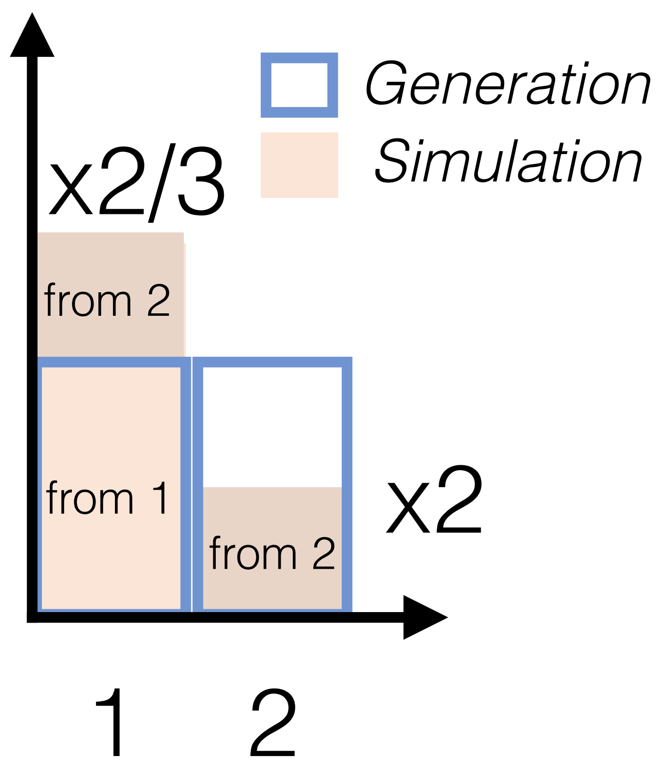


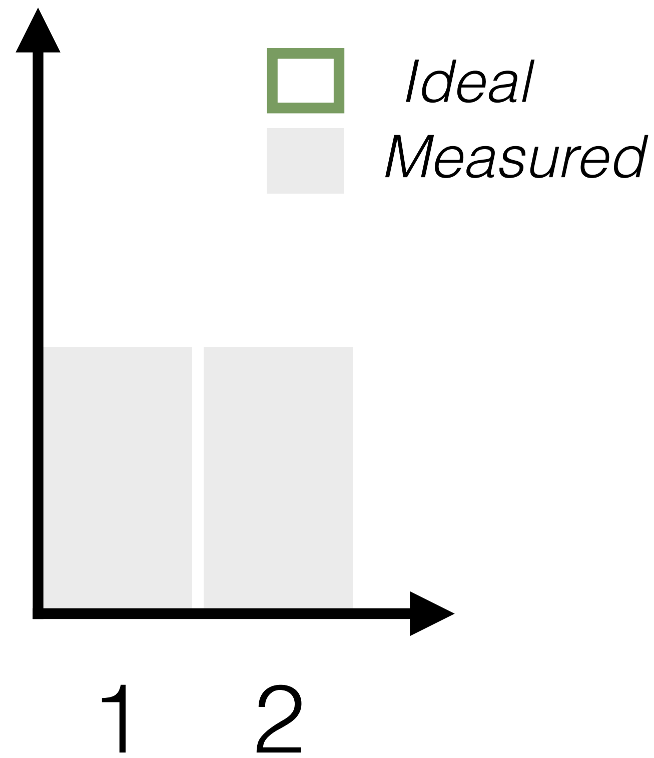




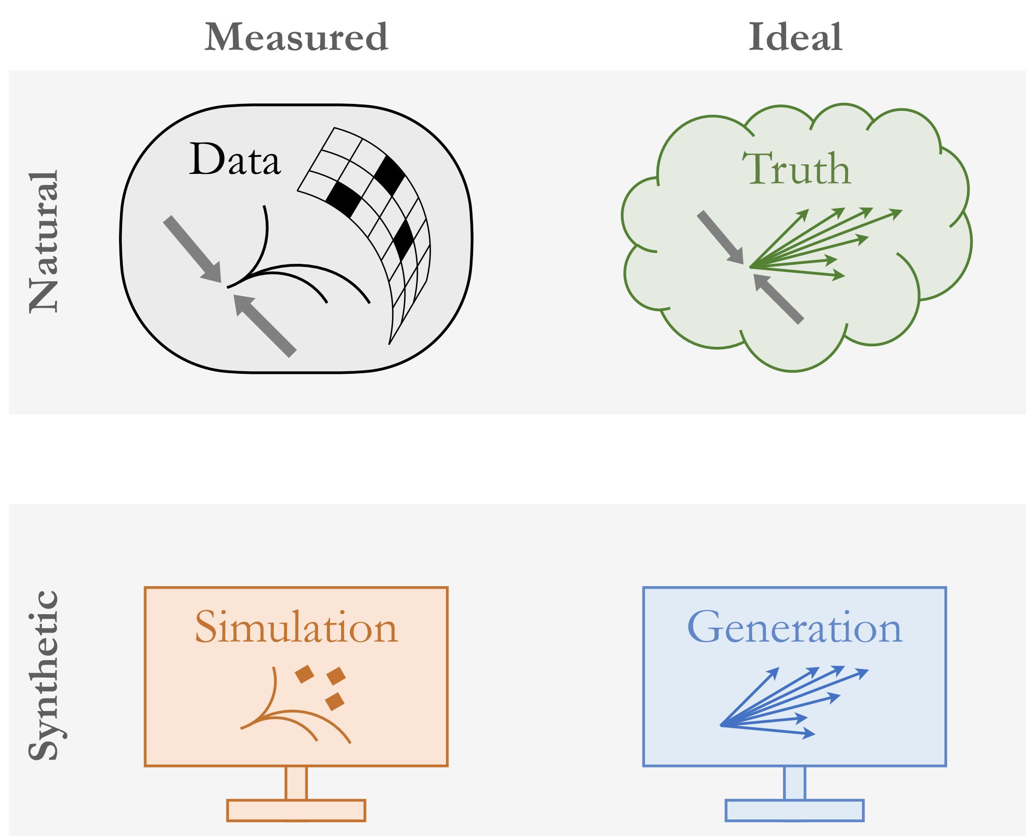
A. Andreassen, P. Komiske, E. Metodiev, BPN, J. Thaler, PRL 124 (2020) 182001

Unfold by iterating: OmniFold





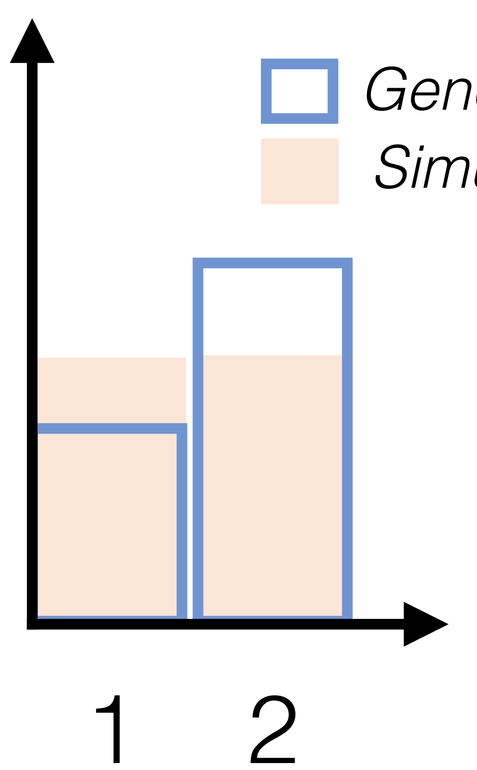
Measured \ Ideal	1	2
2	0%	50%
1	100%	50%

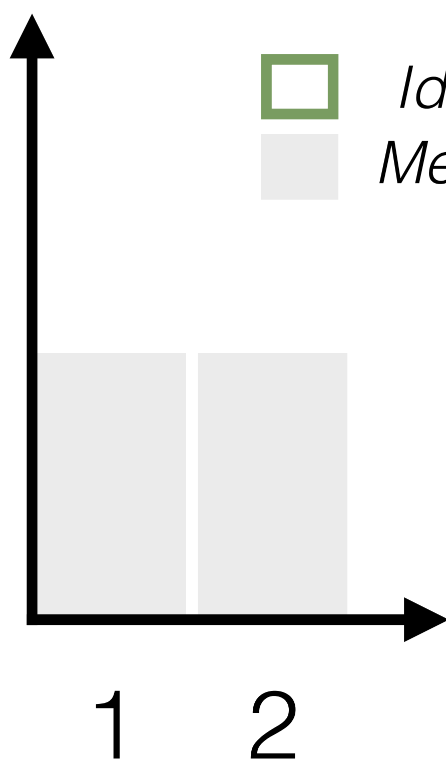


A. Andreassen, P. Komiske, E. Metodiev, BPN, J. Thaler, PRL 124 (2020) 182001

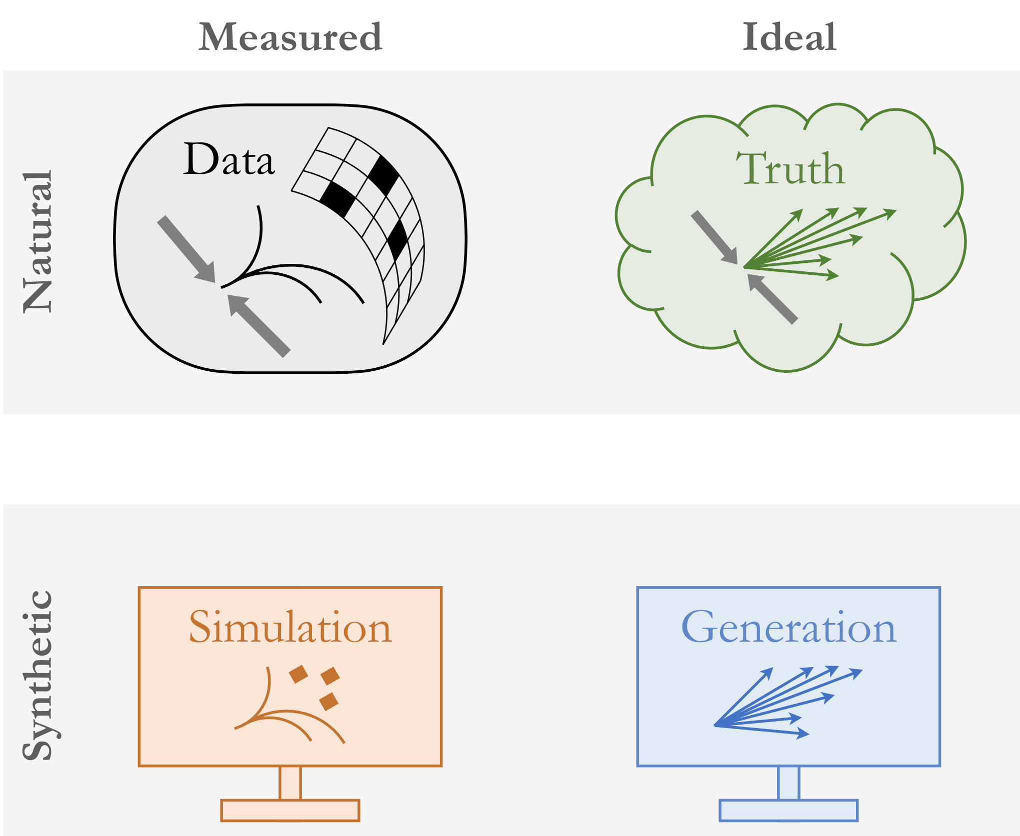
Unfold by iterating: OmniFold

After iteration 1





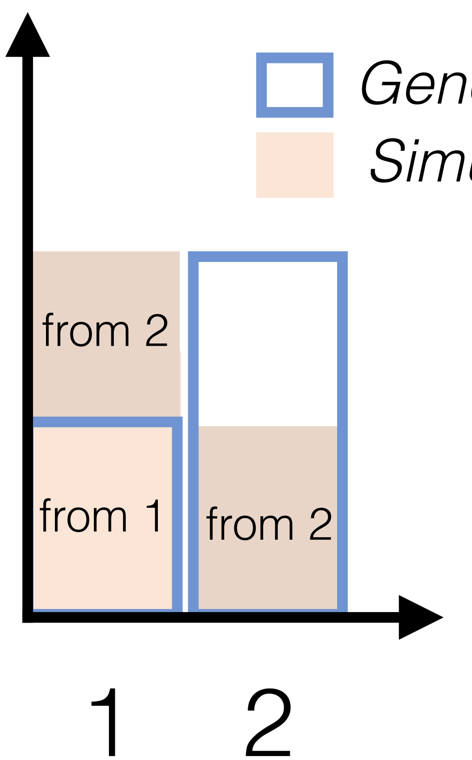
Measured \ Ideal	1	2
2	0%	50%
1	100%	50%

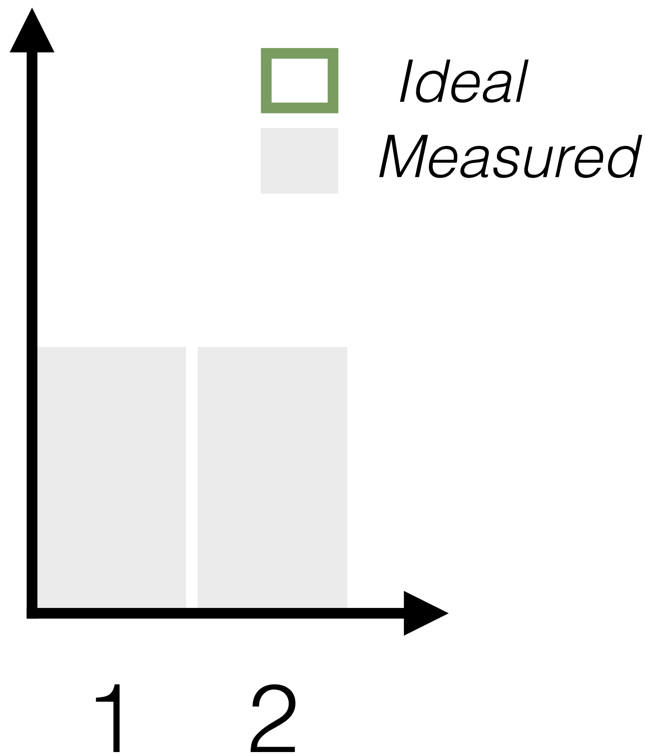


A. Andreassen, P. Komiske, E. Metodiev, BPN, J. Thaler, PRL 124 (2020) 182001

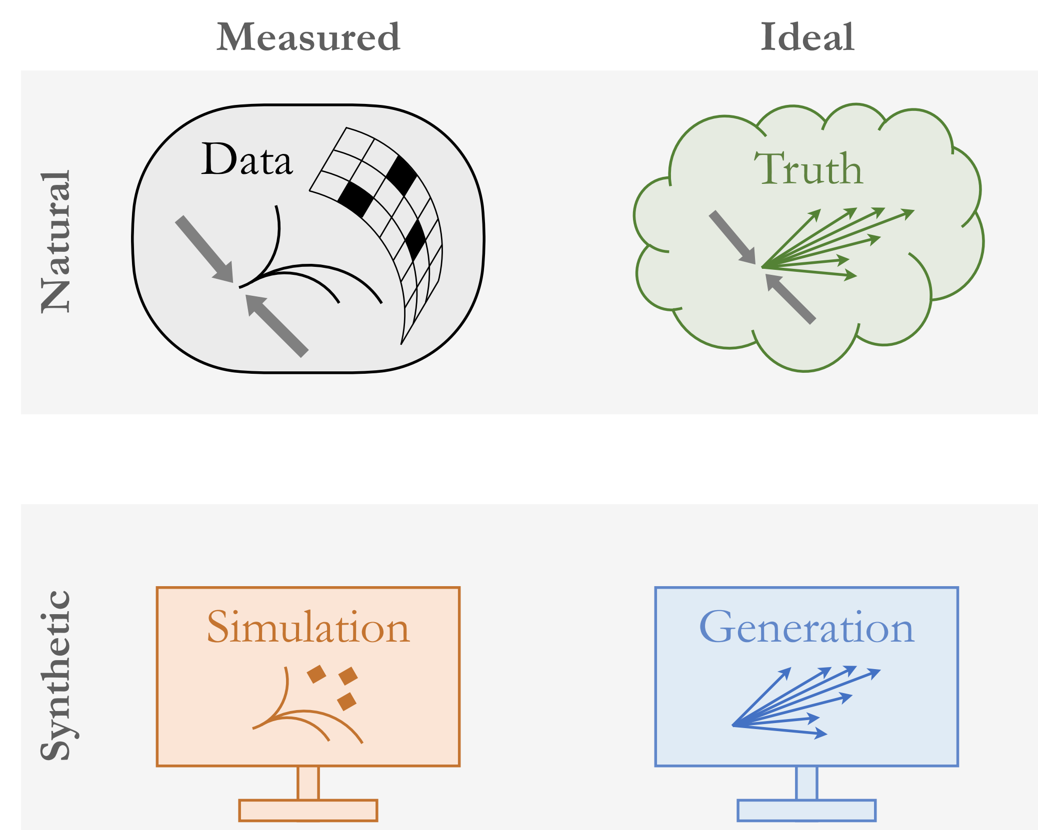
Unfold by iterating: OmniFold

After iteration 1





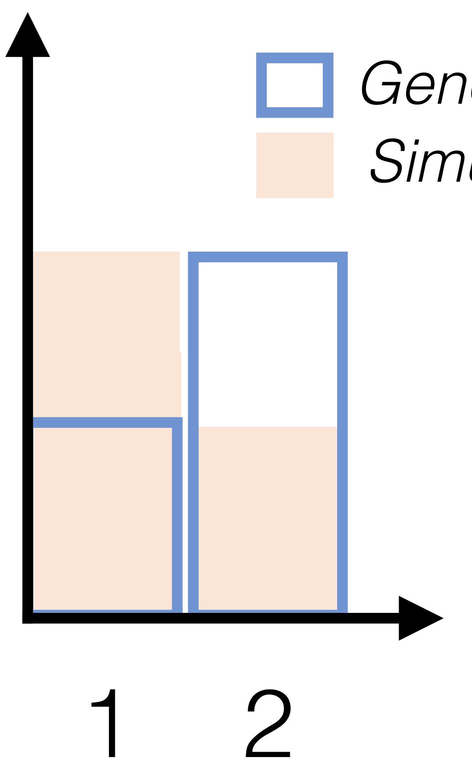
Measured \ Ideal	1	2
2	0%	50%
1	100%	50%

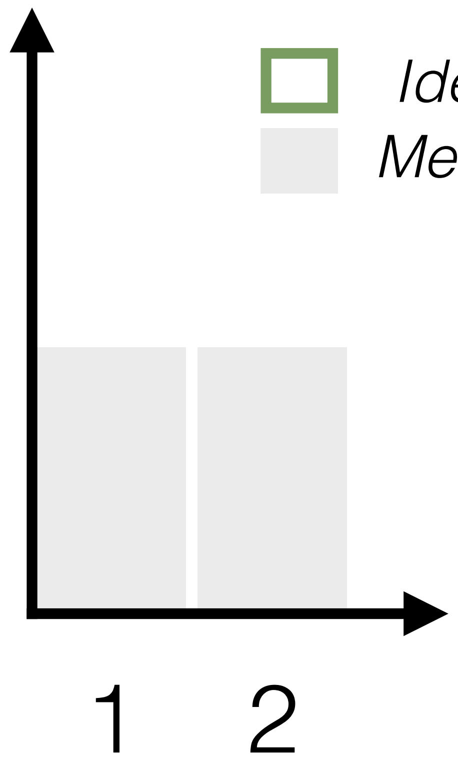


A. Andreassen, P. Komiske, E. Metodiev, BPN, J. Thaler, PRL 124 (2020) 182001

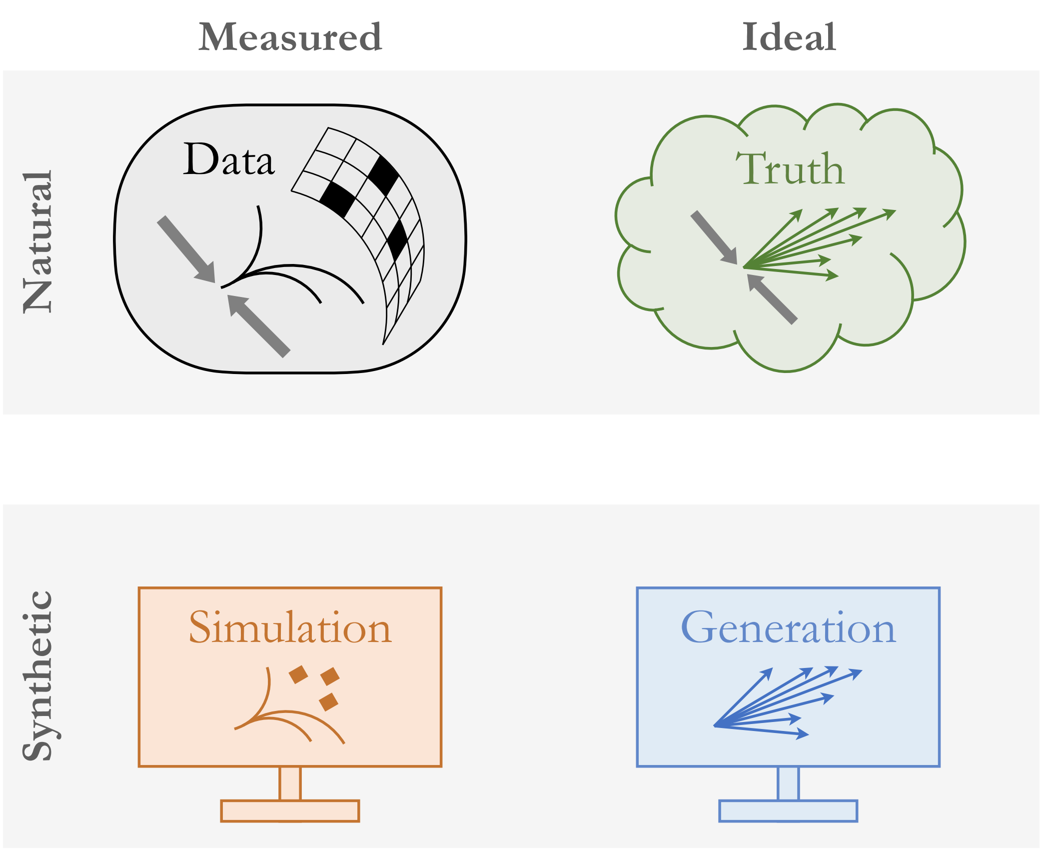
Unfold by iterating: OmniFold

After iteration 1





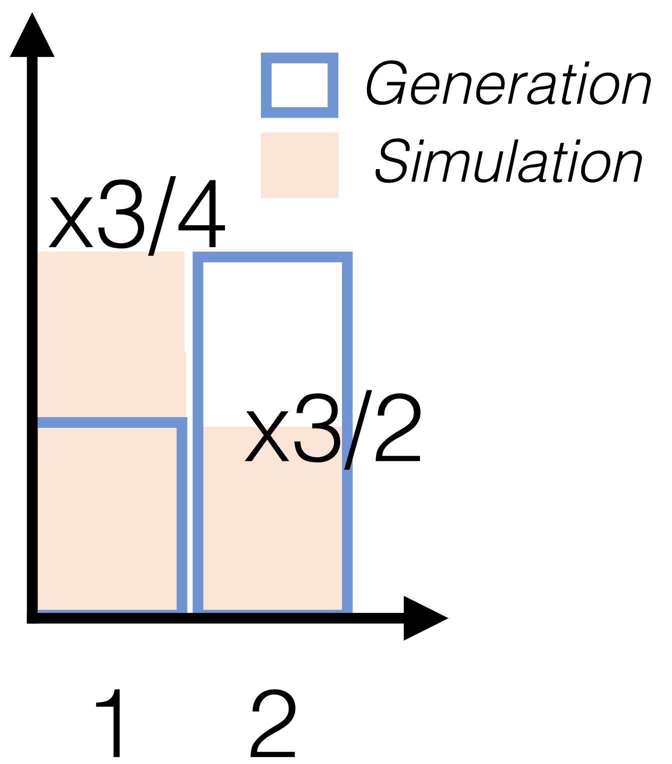
Measured		Ideal 1	Ideal 2
	2	0%	50%
	1	100%	50%

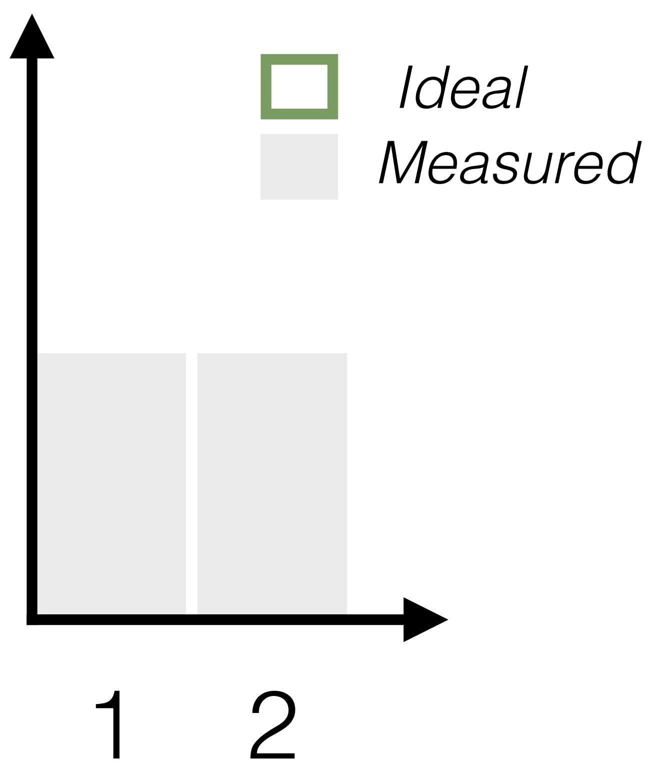


A. Andreassen, P. Komiske, E. Metodiev, BPN, J. Thaler, PRL 124 (2020) 182001

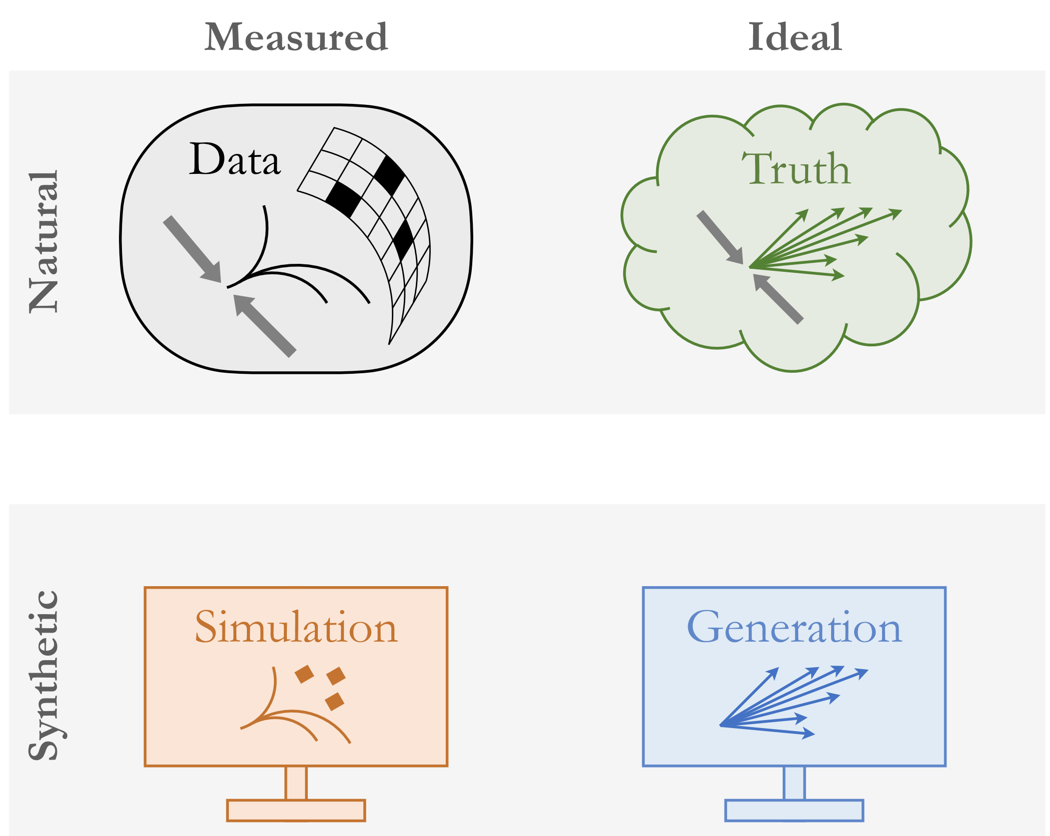
Unfold by iterating: OmniFold

After iteration 1





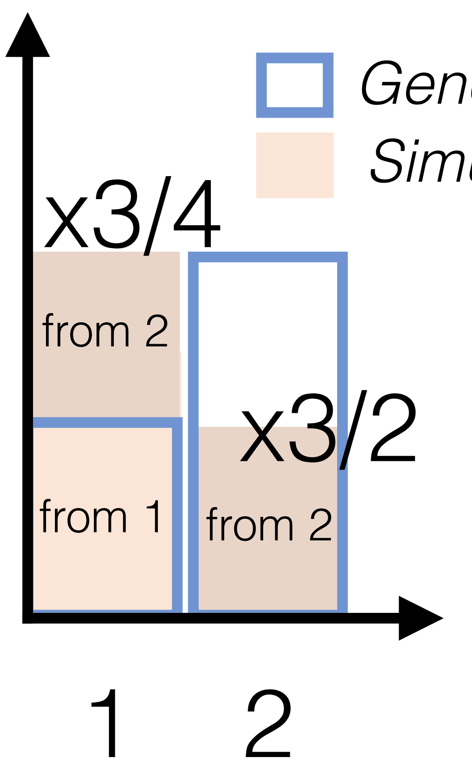
Measured \ Ideal	1	2
2	0%	50%
1	100%	50%

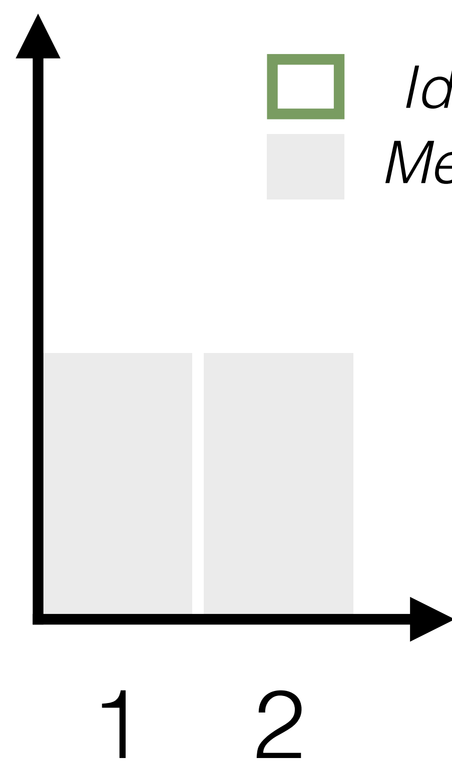


A. Andreassen, P. Komiske, E. Metodiev, BPN, J. Thaler, PRL 124 (2020) 182001

Unfold by iterating: OmniFold

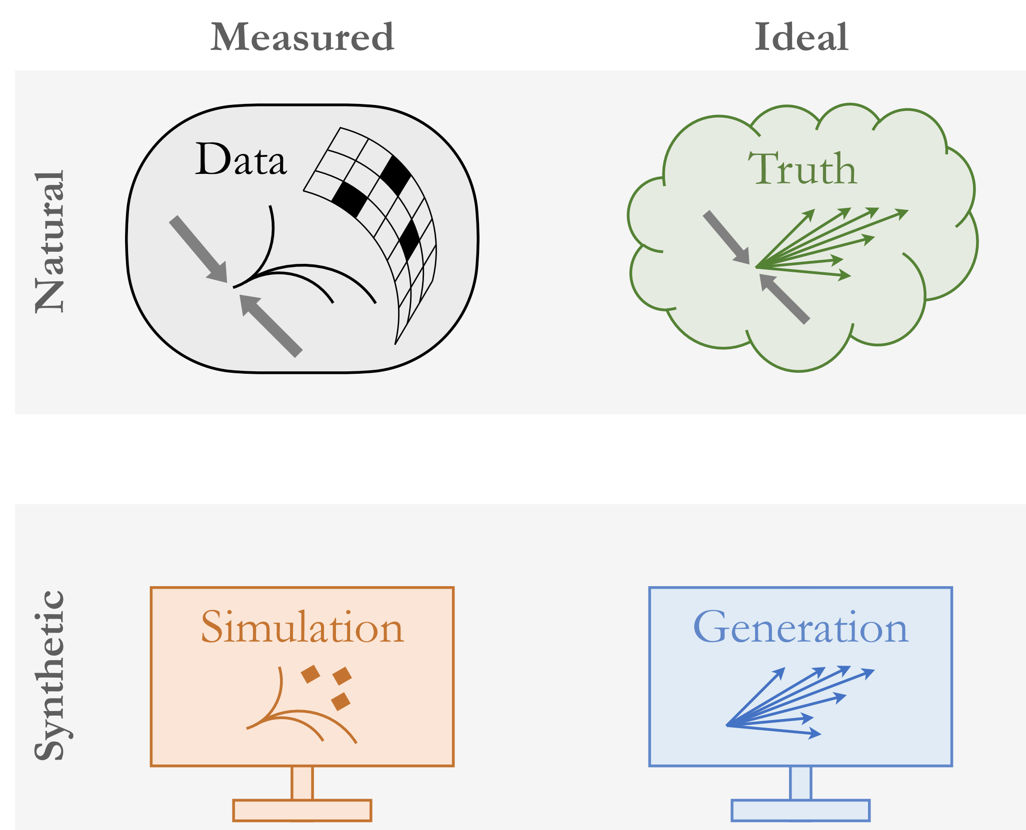
After iteration 1





Measured (rows) vs. Ideal (columns):

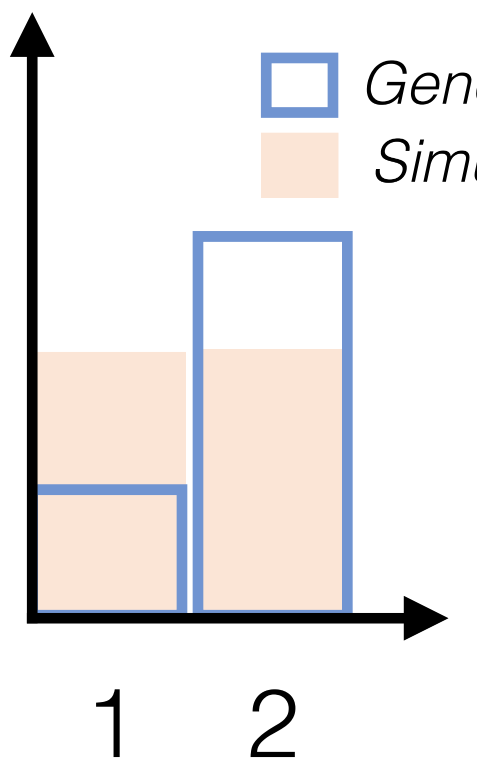
Measured \ Ideal	1	2
2	0%	50%
1	100%	50%

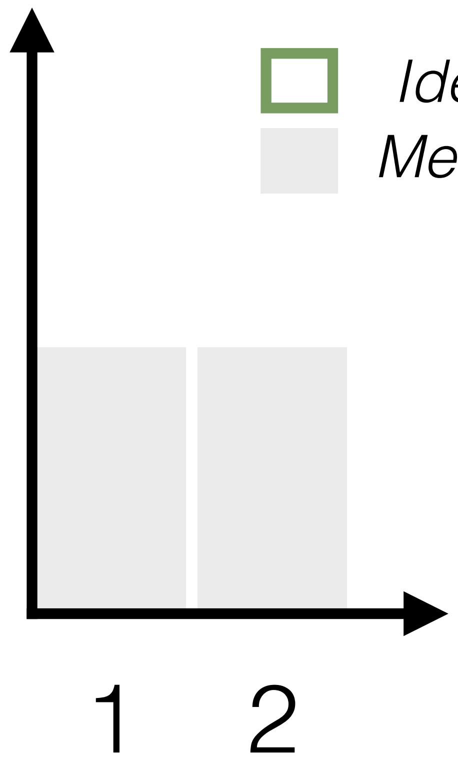


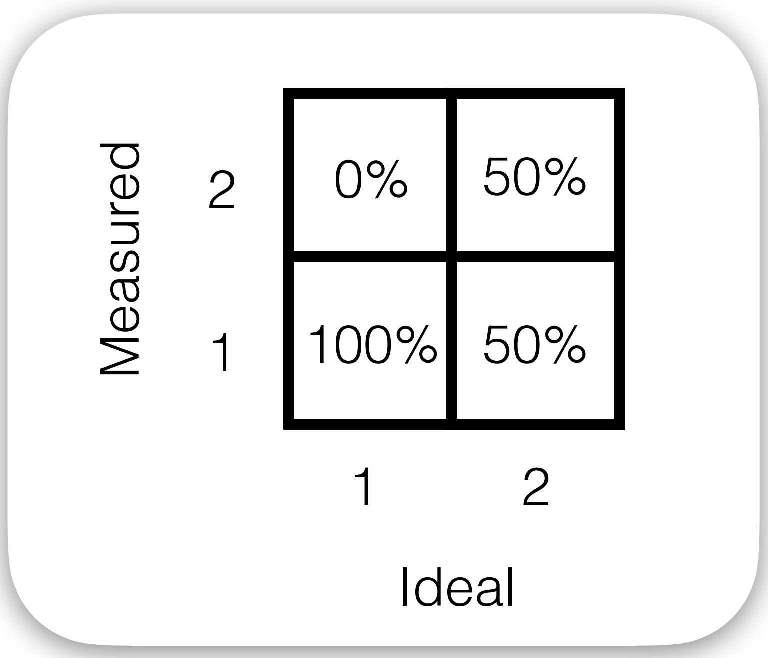
A. Andreassen, P. Komiske, E. Metodiev, BPN, J. Thaler, PRL 124 (2020) 182001

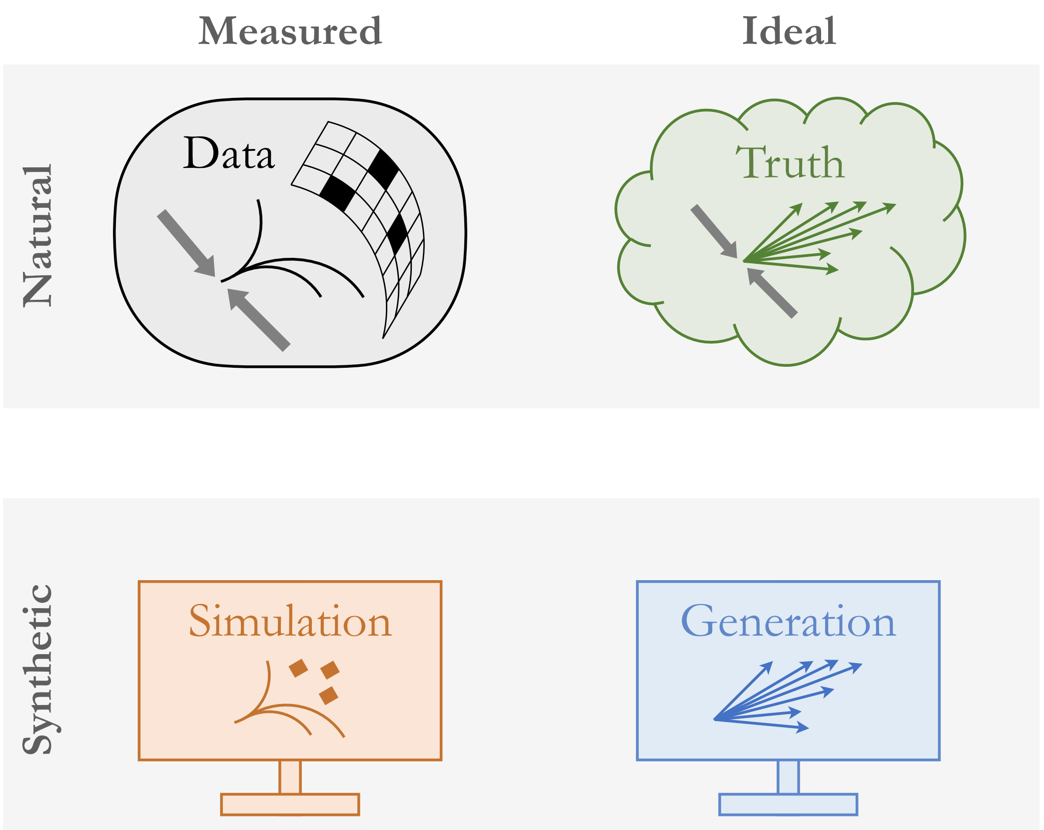
Unfold by iterating: OmniFold

After iteration 2





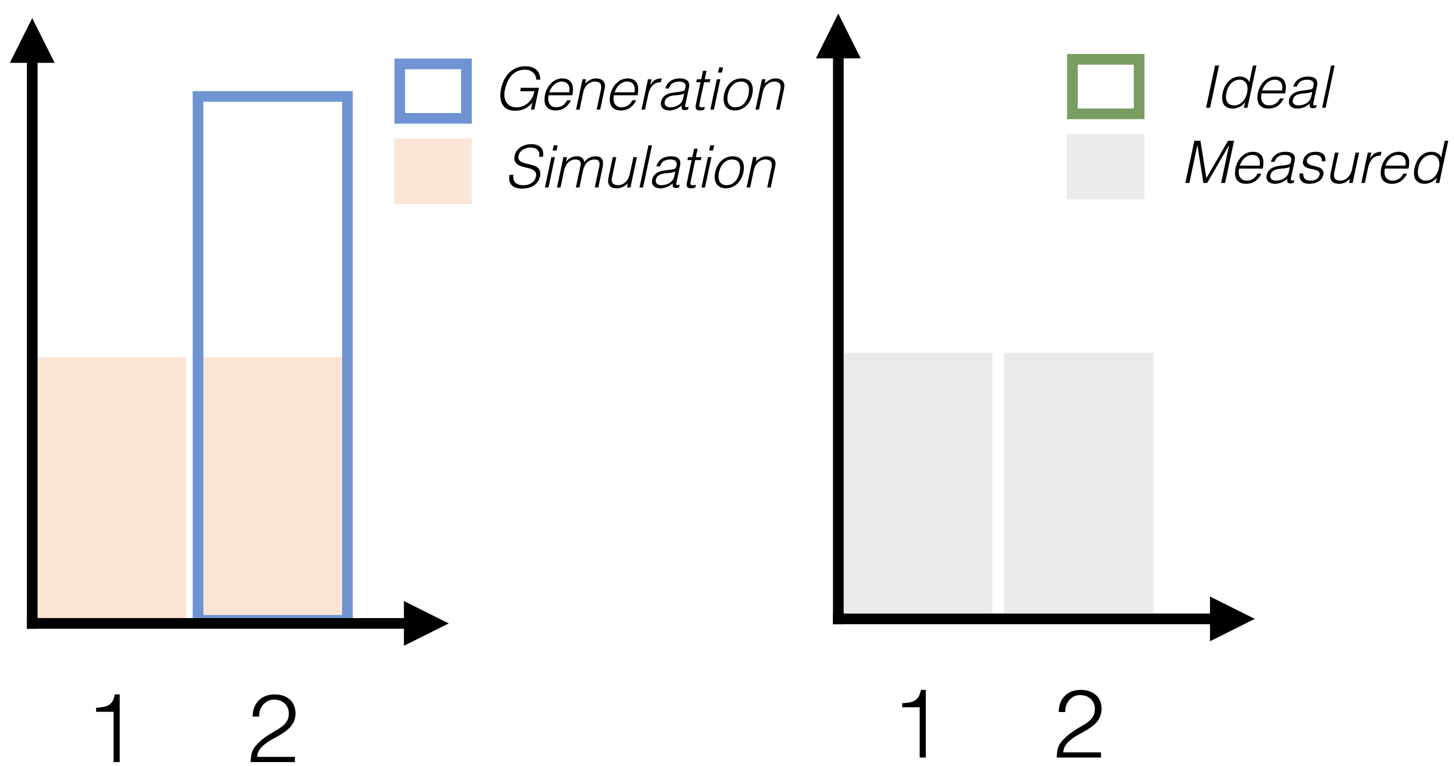




A. Andreassen, P. Komiske, E. Metodiev, BPN, J. Thaler, PRL 124 (2020) 182001

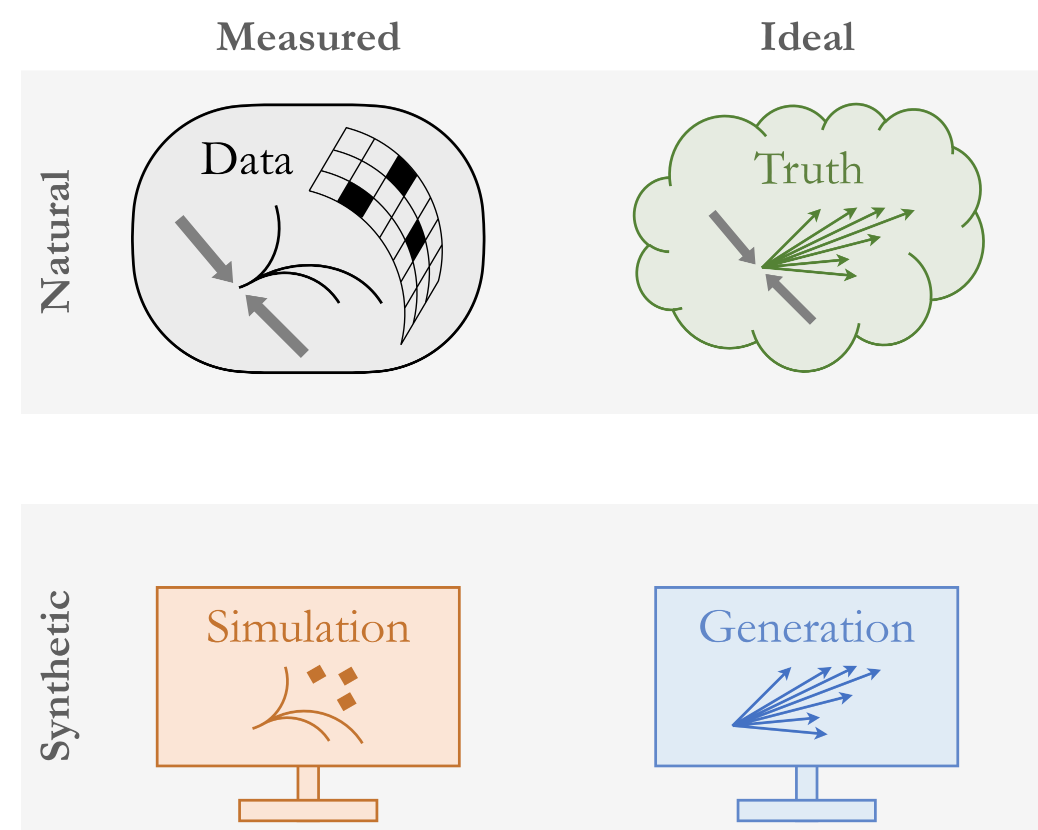
Unfold by iterating: OmniFold

After iteration ∞



N.B. if you just apply p(ideal | measured), *you would have gotten the wrong answer!*

Measured \ Ideal	1	2
2	0%	50%
1	100%	50%



A. Andreassen, P. Komiske, E. Metodiev, BPN, J. Thaler, PRL 124 (2020) 182001

Results

Simultaneous for free!
(binning is for illustration)